

VISUAL QUESTION ANSWERING AND BEYOND

A Dissertation
Presented to
The Academic Faculty

by

Aishwarya Agrawal

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
December 2019

Copyright © 2019 by Aishwarya Agrawal

VISUAL QUESTION ANSWERING AND BEYOND

Approved by:

Dr. Dhruv Batra, Advisor
Georgia Institute of Technology

Dr. C. Lawrence Zitnick
Facebook AI Research

Dr. Devi Parikh
Georgia Institute of Technology

Dr. Oriol Vinyals
Google DeepMind

Dr. James Hays
Georgia Institute of Technology

Date Approved: 20 August 2019

ACKNOWLEDGEMENTS

I would like to thank the following people who played important roles along different dimensions of my life as a PhD student.

First of all, I am deeply thankful to my PhD advisor, Dhruv Batra. When I applied for the PhD program, I was just excited to do computer vision research, but did not have enough knowledge of machine learning. Thank you for accepting me as your student in the first place! Thank you for your constant guidance, support and encouragement since then – be it explaining a concept patiently multiple times till I finally get a hold of it, be it reading and providing elaborate comments on paper / research statement drafts so that it does not just make the current draft better, but also teaches me how to write in general, be it giving me ample freedom and time to explore what I want to do next, be it providing encouraging comments on my successes as well as failures, be it standing up for me! You have always guided me to seek to the best and push beyond what seems achievable. You have cared about me not just as your student but as your kid! Thank you for shaping my career, my life!

Similarly, I am immensely grateful to Devi Parikh, who I not just collaborated with during almost my entire PhD, but also reached out to seek solutions to various problems of my work life (both research related and otherwise). And surprisingly, Devi always had a solution! Your quick and useful replies, no matter how busy you are, your staying up till late night (even when you were not planning to!) to provide feedback on paper drafts, your getting excited when we (your students) are giving talks, your being considerate of everyone's feelings – all these make me admire you a lot! You have not only shaped my research, but also my thinking, my outlook towards life! Thank you for giving me an opportunity to be a part of your life!

I would also like to thank Larry Zitnick for being a wonderful collaborator and mentor. I have always been awed by your data visualization skills (!) and creative thinking in general. Thank you for helping shape my research and for providing useful advice at critical points of my PhD career!

I am also thankful to Oriol Vinyals for engaging in very informative brainstorming sessions with me while I was an intern at DeepMind, for agreeing to be a part of my PhD committee and providing useful feedback on my thesis, and for providing useful advice about life in general!

I would also like to thank James Hays for agreeing to be a part of my PhD committee and providing useful feedback about future directions of my PhD research!

During my PhD, I had the opportunity to intern at a few amazing places in the world. I am grateful to all my internship mentors and collaborators for providing me one of the best experiences of my PhD and for teaching me so much, both about work and life – Margaret Mitchell, Xiaodong He, Aniruddha Kembhavi, Tejas Kulkarni, Mateusz Malinowski, Felix Hill, Ali Eslami, and Jianfeng Gao.

One person who turned out to be like an angel in my life when I started my grad school was my labmate Stanislaw Antol. Stan and I collaborated on two research projects (the first two projects of my PhD). Stan being part of those projects was immensely helpful to me. He helped me with every little thing– from how to run jobs on cluster to how to write good code, from how to setup mechanical turk experiments to how to make sure the incoming data is good. In addition to helping me get settled in a new work environment, he was always helpful outside work too. I still miss his cookies, the hiking trips and Thanksgiving dinner with him. Thank you for all your kindness and support!

I am also immensely thankful to Jiasen Lu, Gordon Christie, Ankit Laddha, Akrit Mohapatra, and Sainandan Ramakrishnan, who I had the pleasure to work closely with and learn a lot from. Thank you for being such wonderful collaborators!

I would like to express my deep gratitude towards my colleagues-cum-friends – Stefan Lee, Ramakrishna Vedantam, Arjun Chandrasekaran, Abhishek Das, Harsh Agrawal, Ramprasaath Selvaraju, and Ashwin Kalyan, who acted as mentors by providing useful advice during critical times, as well as acted as family by filling my life with fun and sweet memories! Thank you for all your support!

I would like to thank all the past and present members of the CVMLP group for making my PhD journey so wonderful and unforgettable – Neelima Chavali, Qing Sun, Michael Cogswell, Xiao Lin, Peng Zhang, Shrenik Lad, Jianwei Yang, Nirbhay Modhe, Samyak Datta, Deshraj Yadav, Viraj Prabhu, Prithvijit Chattopadhyay, Karan Desai, Rishabh Jain, Sanyam Agrawal, Ayush Shrivastava, Mohit Sharma, Purva Tendulkar, Erik Wijmans, Peter Anderson, Zhile Ren, Faruk Ahmed, Clint Solomon, Arijit Ray, Aroma Mahendru, Latha Pemula, Tejas Khot, Khushi Gupta, Avi Singh, Ahmed Osman, Senthil Purushwalkam, Varun Manjunatha, Mayu Sakurada, and Tanmay Batra. It is an honor for me that I got to be a part of the CVMLP group.

I am deeply grateful to my parents Indu Agrawal and Sanjay Agrawal for sacrificing comforts of their life so that they could provide me the best possible education, for letting me follow my dreams even if that meant being thousands of miles away from their only child, for understanding my life choices and being happy about them, for everything! This PhD would not have been possible if you were not this understanding and supportive!

One person who has always believed in me and helped me emotionally, mentally, as well as intellectually is my partner-cum-colleague Yash Goyal. I admire your empathetic nature, I admire your ability to see and show me the positive side of most of the negative-seeming situations of my life. Thank you for being with me during the highs and lows of my life. Thank you for your constant support and encouragement. My PhD would not be a success if it were not for you!

TABLE OF CONTENTS

| | |
|--|------------|
| ACKNOWLEDGEMENTS | iii |
| LIST OF TABLES | ix |
| LIST OF FIGURES | xi |
| SUMMARY | xx |
| I INTRODUCTION | 1 |
| 1.1 Free-form and Open-Ended VQA (chapter 3) | 2 |
| 1.2 Analyzing the Behavior of VQA Models (chapter 4) | 3 |
| 1.3 Overcoming Priors in VQA (chapter 5) | 4 |
| 1.4 Contributions | 5 |
| 1.5 List of Publications | 6 |
| II RELATED WORK | 8 |
| 2.1 Visual Question Answering (VQA) | 8 |
| 2.2 Analyzing the Behavior of VQA Models | 10 |
| 2.3 Overcoming Priors in VQA | 11 |
| III VISUAL QUESTION ANSWERING (VQA) | 14 |
| 3.1 Introduction | 14 |
| 3.2 VQA Dataset Collection | 17 |
| 3.3 VQA Dataset Analysis | 22 |
| 3.3.1 Questions | 23 |
| 3.3.2 Answers | 24 |
| 3.3.3 Commonsense Knowledge | 28 |
| 3.3.4 Captions <i>vs.</i> Questions | 29 |
| 3.4 VQA Baselines and Methods | 30 |
| 3.4.1 Baselines | 30 |
| 3.4.2 Methods | 31 |
| 3.4.3 Results | 34 |

| | | |
|-----------|---|-----------|
| 3.5 | VQA Challenge and Workshop | 41 |
| 3.6 | Conclusion and Discussion | 42 |
| IV | ANALYZING THE BEHAVIOR OF VISUAL QUESTION ANSWERING MODELS | 44 |
| 4.1 | Introduction | 44 |
| 4.2 | Behavior Analyses | 44 |
| 4.2.1 | Generalization to novel instances | 46 |
| 4.2.2 | Complete question understanding | 48 |
| 4.2.3 | Complete image understanding | 50 |
| 4.3 | Conclusion | 52 |
| V | OVERCOMING PRIORS IN VISUAL QUESTION ANSWERING | 54 |
| 5.1 | Visual Question Answering under Changing Priors (VQA-CP) | 54 |
| 5.1.1 | Introduction | 54 |
| 5.1.2 | VQA-CP : Dataset Creation and Analysis | 55 |
| 5.1.3 | Benchmarking VQA Models on VQA-CP | 58 |
| 5.1.4 | Conclusion | 60 |
| 5.2 | Grounded Visual Question Answering (GVQA) | 60 |
| 5.2.1 | Introduction | 60 |
| 5.2.2 | GVQA model | 61 |
| 5.2.3 | Experiments on VQA-CP v1 and VQA-CP v2 | 66 |
| 5.2.4 | Role of GVQA Components | 67 |
| 5.2.5 | Experiments on VQA v1 and VQA v2 | 69 |
| 5.2.6 | Transparency | 71 |
| 5.2.7 | Conclusion | 73 |
| 5.3 | Adversarial Regularization for Visual Question Answering | 73 |
| 5.3.1 | Introduction | 73 |
| 5.3.2 | Reducing Language Bias Through Adversarial Regularization | 75 |
| 5.3.3 | Experiments | 80 |

| | | |
|------------|--|-----|
| 5.3.4 | Results | 82 |
| 5.3.5 | Conclusion | 87 |
| VI | CONCLUSION | 88 |
| APPENDIX A | — APPENDIX FOR VISUAL QUESTION ANSWER- ING | 93 |
| APPENDIX B | — APPENDIX FOR ANALYZING THE BEHAV- IOR OF VQA MODELS | 111 |
| APPENDIX C | — APPENDIX FOR OVERCOMING PRIORS IN VQA | 124 |
| REFERENCES | | 134 |

LIST OF TABLES

| | | |
|---|---|----|
| 1 | Test-standard accuracy of human subjects when asked to answer the question without seeing the image (Question), seeing just a caption of the image and not the image itself (Question + Caption), and seeing the image (Question + Image). Results are shown for all questions, “yes/no” & “number” questions, and other questions that are neither answered “yes/no” nor number. All answers are free-form and not multiple-choice. *These accuracies are evaluated on a subset of 3K train questions (1K images). | 30 |
| 2 | Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val). See text for details. | 34 |
| 3 | Open-ended test-dev results for different question types on real images (Q+C is reported on val). Machine performance is reported using the bag-of-words representation for questions. Questions types are determined by the one or two words that start the question. The percentage of questions for each type is shown in parentheses. Last and second last columns respectively show the average human age and average degree of commonsense required to answer the questions (as reported by AMT workers), respectively. See text for details. | 36 |
| 4 | Accuracy of ablated versions of our best model (deeper LSTM Q + norm I) for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image. See text for details. | 41 |
| 5 | Test-standard accuracy of our best model (deeper LSTM Q + norm I) compared to test-standard accuracies of other entries for the open-ended and multiple-choice tasks in the respective VQA Real Image Challenge leaderboards (as of October 28, 2016). | 42 |
| 6 | We compare the performance of existing VQA models on VQA-CP v1 test splits (when trained on VQA-CP v1 train splits) to their performance on VQA v1 val splits (when trained on VQA v1 train splits). We find that the performance of all tested existing models degrades significantly in the new Changing Priors setting compared to the original VQA setting. | 58 |

| | | |
|----|---|-----|
| 7 | We compare the performance of existing VQA models on VQA-CP v2 test splits (when trained on VQA-CP v2 train splits) to their performance on VQA v2 val splits (when trained on VQA v2 train splits). We find that the performance of all tested existing models degrades significantly in the new Changing Priors setting compared to the original VQA setting. | 59 |
| 8 | Performance of GVQA (our model) compared to SAN on VQA-CP datasets. GVQA consistently outperforms SAN. | 66 |
| 9 | Experimental results when each component in GVQA (denoted by “-<component>”) is replaced with its corresponding traditional counterpart (denoted by “+ <traditional counterpart>”). | 69 |
| 10 | Results of GVQA and SAN on VQA v1 and VQA v2 when trained on the corresponding train splits. | 69 |
| 11 | Performance on VQA-CP v2 test and VQA v2 val . We significantly improve the accuracy of base models and achieve state-of-the-art performance on the VQA-CP dataset. | 82 |
| 12 | Performance on VQA-CP v1 test and VQA v1 val | 83 |
| 13 | For each of the two datasets, real and abstract, first two rows are the human accuracies for multiple-choice questions when subjects were shown both the image and the question. Majority vote means we consider the answer picked by majority of the three subjects to be the predicted answer by humans and compute accuracy of that answer for each question. Average means we compute the accuracy of each of the answers picked by the subjects and record their average for each question. The last row is the inter-human agreement for open-ended answers task when subjects were shown both the image and the question. All accuracies are evaluated on a random subset of 3000 questions. | 100 |
| 14 | Performance of SAN - $Q_{full} + Q_{main}$ compared to SAN and GVQA (our model) on VQA-CP v2 dataset. GVQA outperforms both SAN and SAN - $Q_{full} + Q_{main}$ | 129 |
| 15 | Results of GVQA, GVQA - VCC_{loss} and SAN on VQA v1 val split when trained on the VQA v1 train split. Please see text for more details. | 129 |
| 16 | Results of GVQA, GVQA - VCC_{loss} and SAN on VQA v2 val split when trained on the VQA v2 train split. Please see text for more details. | 130 |

LIST OF FIGURES

| | | |
|---|--|----|
| 1 | Examples of free-form, open-ended questions in our VQA dataset. . . | 3 |
| 2 | This figure illustrates outputs from a baseline model (SAN [495]) and the proposed model (GVQA [12]). For the given test questions, SAN predicts the prior answers from the training data for the respective question types, resulting in incorrect answers. However, GVQA, being more visually grounded than SAN, correctly answers the test questions. | 4 |
| 3 | Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions. | 15 |
| 4 | Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples. | 17 |
| 5 | Distribution of questions by their first four words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show. | 23 |
| 6 | Percentage of questions with different word lengths for real images and abstract scenes. | 24 |
| 7 | Distribution of answers per question type for a random sample of 60K questions for real images when subjects provide answers when given the image (top) and when not given the image (bottom). | 25 |
| 8 | Number of questions per average confidence score (0 = not confident, 1 = confident) for real images and abstract scenes (black lines). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same (color bars). | 27 |
| 9 | 1214.5=13.613.6Example questions judged by Mturk workers to be answerable by different age groups. The percentage of questions falling into each age group is shown in parentheses. | 27 |

| | | |
|----|---|----|
| 10 | Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [408] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers. | 32 |
| 11 | Pr(system is correct answer) for 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). System refers to our best model (deeper LSTM Q + norm I). | 38 |
| 12 | Pr(answer system is correct) for 50 most frequently predicted answers on the VQA validation set (plot is sorted by prediction frequency, not accuracy). System refers to our best model (deeper LSTM Q + norm I). | 38 |
| 13 | Pr(system is correct age of question) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I). | 39 |
| 14 | Pr(age of question system is correct) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I). | 40 |
| 15 | Leaderboard showing test-standard accuracies for VQA Real Image Challenge (Open-Ended) on left and leaderboard showing test-standard accuracies for VQA Real Image Challenge (Multiple-Choice) on right (snapshot from October 28, 2016). | 43 |
| 16 | Examples from test set where the CNN+LSTM model makes mistakes and their corresponding nearest neighbor training instances. See appendix for more examples. | 48 |
| 17 | X-axis shows length of partial question (in %) fed as input. Y-axis shows percentage of questions for which responses of these partial questions are the same as full questions and VQA accuracy of partial questions. | 49 |
| 18 | Examples where the CNN+LSTM model does not change its answer after first few question words. On doing so, it is correct for some cases (the extreme left example) and incorrect for other cases (the remaining three examples). See appendix for more examples. | 50 |
| 19 | Percentage of questions for which responses remain same (compared to entire question) as a function of POS tags dropped from the question. | 51 |

| | | |
|----|---|----|
| 20 | Histogram of percentage of images for which model produces same answer for a given question and its comparison with test accuracy. The cumulative plot shows the % of questions for which model produces same answer for <i>atleast x</i> % of images. | 52 |
| 21 | Examples where the predicted answers do not change across images for a given question. See appendix for more examples. | 52 |
| 22 | Distribution of answers per question type vary significantly between VQA-CP v1 train (left) and test (right) splits. For instance, ‘ <i>white</i> ’ and ‘ <i>red</i> ’ are commonly seen answers in train for ‘ <i>What color</i> ’, where as ‘ <i>black</i> ’ is the most frequent answer in test. These have been computed for a random sample of 60K questions. | 57 |
| 23 | The proposed Grounded Visual Question Answering (GVQA) model. | 62 |
| 24 | Qualitative examples from GVQA. Left: We show top three answer cluster predictions (along with random concepts from each cluster) by ACP. Corresponding to each cluster predicted by ACP, we show the top visual concept predicted by VCC. Given these ACP and VCC predictions, the Answer Predictor (AP) predicts the correct answer ‘ <i>baseball</i> ’. Right: Smiling is the concept extracted by the CE whose visual presence in VCC’s predictions is verified by the Visual Verifier, resulting in ‘ <i>yes</i> ’ as the final answer. | 71 |
| 25 | Left: GVQA’s prediction (‘ <i>green</i> ’) can be explained as follows – ACP predicts that the answer should be a <i>color</i> . Of the various visual concepts predicted by VCC, the only concept that is about color is <i>green</i> . Hence, GVQA’s output is ‘ <i>green</i> ’. SAN incorrectly predicts ‘ <i>yellow</i> ’. SAN’s architecture doesn’t facilitate producing an explanation of why it predicted what it predicted, unlike GVQA. Right: Both GVQA and SAN incorrectly answer the question. GVQA is incorrect perhaps because VCC predicts ‘ <i>black</i> ’, instead of ‘ <i>gray</i> ’. In order to dig further into why VCC’s prediction is incorrect, we can look at the attention map (in appendix), which shows that the attention is on the pants for the person’s left leg, but on the socks (black in color) for the person’s right leg. So, perhaps, VCC’s “black” prediction is based on the attention on the person’s right leg. | 72 |

| | | |
|----|--|----|
| 26 | Given an arbitrary base VQA model (A), we introduce two regularizers. First, we build a question-only adversary (B) that takes the question embedding \mathbf{q}_i from the VQA model and is trained to output the correct answer from this information alone. For this network to succeed, \mathbf{q}_i must capture language biases from the dataset – the same biases that lead the base VQA model to ignore visual content. To reduce these biases, we set the base VQA model and the question-only adversary against each other, with the base VQA network modifying its question embedding to reduce question-only performance (shown here as gradient negation of the question-only model loss) Further, the question-only model allows estimation of the change in answer confidence given image (C), which we maximize explicitly. | 75 |
| 27 | Maximizing difference of entropies (DoE) along with the question-only adversarial regularization for the SAN model, not only improves results on changing priors, but also stabilizes training. | 85 |
| 28 | Answer distribution for SAN+Q-Adv+DoE mimic the prior less for questions with high language bias. | 85 |
| 29 | The table-top setup and an example dialog from the SHRDLU [477] project (studied by Terry Winograd in 1972). | 91 |
| 30 | Given an instruction (<i>‘There is a small sphere’</i>), the the task for an agent is to execute actions to create scenes that are consistent with the given instruction (i.e., each such scene consists of a small sphere). . . | 92 |
| 31 | Proportions of spatial prepositions in the captions and question & answers for real images (left) and abstract scenes (right). | 95 |
| 32 | Venn-style word clouds [100] for nouns with size indicating the normalized count for real images. | 95 |
| 33 | Venn-style word clouds [100] for verbs with size indicating the normalized count for real images. | 96 |
| 34 | Venn-style word clouds [100] for adjectives with size indicating the normalized count for real images. | 96 |
| 35 | Venn-style word clouds [100] for nouns with size indicating the normalized count for abstract scenes. | 97 |
| 36 | Venn-style word clouds [100] for verbs with size indicating the normalized count for abstract scenes. | 97 |
| 37 | Venn-style word clouds [100] for adjectives with size indicating the normalized count for abstract scenes. | 97 |

| | | |
|----|--|-----|
| 38 | Distribution of questions starting with “What is” by their first five words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show. | 98 |
| 39 | Distribution of answers for questions starting with “What is” for a random sample of 60K questions for real images (top) and all questions for abstract scenes (bottom). Each column corresponds to questions ending in different words, such as “doing?”, “on?”, <i>etc.</i> | 99 |
| 40 | Left: A small subset of the objects present in the abstract scene dataset. Right: The AMT interface for collecting abstract scenes. The light green circles indicate where users can select to manipulate a person’s pose. Different objects may be added to the scene using the folders to the right. | 102 |
| 41 | Our AMT interface for collecting the third question for an image, when subjects were shown previous questions that were collected and were asked to ask a question different from previous questions. | 103 |
| 42 | The AMT interface used to collect answers to a question when subjects were shown the image while answering the question. | 104 |
| 43 | The AMT interface used to collect answers to a question when subjects were not shown the image while answering the question using only commonsense to collect the plausible, but incorrect, multiple-choice answers. | 104 |
| 44 | Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset. | 108 |
| 45 | Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the abstract scene dataset. | 109 |
| 46 | Random examples of multiple-choice questions for numerous representative examples of the real and abstract scene dataset. | 110 |
| 47 | Test accuracy vs. average distance of the test points from k-NN training points for the CNN+LSTM model. | 112 |
| 48 | Test QI pairs for which the CNN+LSTM model produces the correct response and their nearest neighbor QI pairs from training set. | 115 |

| | | |
|----|---|-----|
| 49 | Test QI pairs for which the CNN+LSTM model produces incorrect response and their nearest neighbor QI pairs from training set. | 116 |
| 50 | X-axis shows length of partial “yes/no” question (in %) fed as input. Y-axis shows percentage of “yes/no” questions for which responses of these partial “yes/no” questions are the same as full “yes/no” questions and VQA accuracy of partial “yes/no” questions. | 117 |
| 51 | X-axis shows length of partial “number” question (in %) fed as input. Y-axis shows percentage of “number” questions for which responses of these partial “number” questions are the same as full “number” questions and VQA accuracy of partial “number” questions. | 117 |
| 52 | X-axis shows length of partial “other” question (in %) fed as input. Y-axis shows percentage of “other” questions for which responses of these partial “other” questions are the same as full “other” questions and VQA accuracy of partial “other” questions. | 118 |
| 53 | Percentage of “yes/no” questions for which responses remain same (compared to entire “yes/no” question) as a function of POS tags dropped from the “yes/no” question. | 118 |
| 54 | Percentage of “number” questions for which responses remain same (compared to entire “number” question) as a function of POS tags dropped from the “number” question. | 119 |
| 55 | Percentage of “other” questions for which responses remain same (compared to entire “other” question) as a function of POS tags dropped from the “other” question. | 119 |
| 56 | Examples where the CNN+LSTM model converges on a predicted answer without listening to the entire question. | 120 |
| 57 | Histogram of percentage of images for which model produces same answer for a given “yes/no” question. The cumulative plot shows the % of “yes/no” questions for which model produces same answer for <i>atleast</i> x % of images. | 121 |
| 58 | Histogram of percentage of images for which model produces same answer for a given “number” question. The cumulative plot shows the % of “number” questions for which model produces same answer for <i>atleast</i> x % of images. | 121 |
| 59 | Histogram of percentage of images for which model produces same answer for a given “other” question. The cumulative plot shows the % of “other” questions for which model produces same answer for <i>atleast</i> x % of images. | 122 |

| | | |
|----|---|-----|
| 60 | Examples where the CNN+LSTM model produces the same answer for atleast half the images for each of the questions shown above. “Q” denotes the question for which model produces same response for atleast half the images, “A” denotes the answer predicted by the model (which is same for atleast half the images), “Number of Images” denotes the number of images for which the question is repeated in the VQA validation set and “Average Accuracy” is the VQA accuracy for these QI pairs (with same question but different images). | 123 |
| 61 | Distribution of answers per question type vary significantly between VQA-CP v2 train (left) and test (right) splits. For instance, ‘ <i>white</i> ’ and ‘ <i>black</i> ’ are commonly seen answers in train for ‘ <i>What color</i> ’, where as ‘ <i>red</i> ’ is the most frequent answer in test. These have been computed for a random sample of 60K questions. | 125 |
| 62 | Performance of SAN and GVQA for different VQA-CP v2 splits. GVQA consistently outperforms SAN across all splits. | 128 |
| 63 | VCC’s attention map for the example shown in Fig. 25 (right) | 131 |
| 64 | Transparency of GVQA. For each of the above examples, GVQA’s intermediate predictions can help explain why it predicted what it predicted. Top-left: VCC predicts the following visual concepts – blue, person, skiing and jacket. ACP predicts the cluster corresponding to colors. Finally, GVQA predicts ‘ <i>blue</i> ’ as the answer. So, we can see why GVQA predicts ‘ <i>blue</i> ’ – because, of all the visual concepts predicted by VCC, only ‘ <i>blue</i> ’ represents a color. Looking at the attention maps can further indicate why GVQA is “seeing” blue (because it is “looking” at the jacket as well, unlike SAN which is only “looking” at the pants). SAN’s prediction is ‘ <i>orange</i> ’ and unlike GVQA, SAN’s architecture does not facilitate producing such an explanation, which makes it difficult to understand why it is saying what it is saying. Top-right: Both GVQA and SAN are “looking” at the regions covered with snow, but GVQA correctly predicts ‘ <i>winter</i> ’, whereas SAN incorrectly predicts ‘ <i>summer</i> ’ which is unclear why. Bottom-left: The Concept Extractor (CE) predicts ‘ <i>happy</i> ’ whose visual presence is verified by VCC which is “looking” at the region corresponding to the kid’s face. Bottom-right: GVQA focuses on a larger part of the scene and correctly recognizes it as ‘ <i>bathroom</i> ’. | 132 |

65 **Transparency of GVQA.** For the above examples, both GVQA and SAN incorrectly answer the question. However, GVQA’s intermediate predictions can help explain why it is incorrect. **Top-left:** For GVQA, VCC’s predictions indicate that it is perhaps “looking” at the field, which can be further verified by the attention map. SAN’s attention map suggests that it is “looking” at the ball but still does not explain why it is predicting ‘*soccer*’. Perhaps, it is confusing the ball with a soccer ball. **Top-right:** The attention maps from GVQA and SAN look similar to each other. However, looking at ACP’s and VCC’s prediction (for GVQA) suggest that it is indeed “seeing” ‘*pasta*’ (the correct answer), but still predicting ‘*carrots*’ because the ACP is incorrectly predicting the cluster corresponding to vegetables instead of the cluster corresponding to pasta. **Bottom:** GVQA is “looking” at the smartphone (unlike SAN), but yet incorrectly answers ‘*no*’, because the VCC does not recognize the phone as a smartphone. It however correctly predicts ‘*phone*’, ‘*electronic*’, ‘*black*’ and ‘*right*’. 133

Thesis Statement

Existing deep visual question answering models tend to rely on language correlations, but can be trained to resist these correlations via appropriate inductive biases and objective functions.

SUMMARY

In this dissertation, I propose and study a multi-modal Artificial Intelligence (AI) task called Visual Question Answering (VQA) – given an image and a natural language question about the image (e.g., ‘*What kind of store is this?*’, ‘*Is it safe to cross the street?*’), the machine’s task is to automatically produce an accurate natural language answer (‘*bakery*’, ‘*yes*’). Applications of VQA include – aiding visually impaired users in understanding their surroundings, aiding analysts in examining large quantities of surveillance data, teaching children through interactive demos, interacting with personal AI assistants, and making visual social media content more accessible.

Specifically, I study the following – 1) how to create a large-scale dataset and define evaluation metrics for free-form and open-ended VQA, 2) how to develop techniques for characterizing the behavior of VQA models, and 3) how to build VQA models that are less driven by language biases in training data and are more visually grounded, by proposing – a) a new evaluation protocol, b) a new model architecture, and c) a novel objective function.

Most of my past work has been towards building agents that can ‘*see*’ and ‘*talk*’. However, for a lot of practical applications (e.g., physical agents navigating inside our houses executing natural language commands) we need agents that can not only ‘*see*’ and ‘*talk*’ but can also take actions. In chapter 6, I present future directions towards generalizing vision and language agents to be able to take actions.

CHAPTER I

INTRODUCTION

One of the goals of Artificial Intelligence (AI) [441] is to develop systems that can ‘see’ (i.e. understand the contents of an image: who, what, where, doing what?) and ‘talk’ (i.e. communicate the understanding to humans in free-form natural language). Applications of such systems include:

- Aiding visually impaired users in understanding their surroundings [51] (Human: ‘*What is on the shelf above the microwave?*’, AI: ‘*Canned containers*’),
- Aiding analysts in making decisions based on large quantities of surveillance data (Human: ‘*What kind of car did the man in red shirt leave in?*’, AI: ‘*Blue Toyota Prius*’),
- Teaching children through interactive demos (Kid: ‘*What animal is that?*’, AI: ‘*That is Dall Sheep. You can find those in Alaska.*’),
- Interacting with personal AI assistants (such as Alexa, Siri) (Human: ‘*Is my laptop in my bedroom upstairs?*’, AI: ‘*Yes*’, Human: ‘*Is the charger plugged in?*’),
- Making visual social media content more accessible (AI: ‘*Your friend Bob just uploaded a picture from his Hawaii trip*’, Human: ‘*Great, is he at the beach?*’, AI: ‘*No, on a mountain*’).

As a first step towards building machines that can convey their understanding of visual content via natural language, in this dissertation, I introduce and study open-ended and free-form Visual Question Answering (VQA) [27, 18] – Given an image and a natural language question about the image (e.g., ‘*What kind of store is this?*’, ‘*How many people are waiting in the queue?*’, ‘*Is it safe to cross the street?*’),

the machine’s task is to automatically produce an accurate natural language answer (*‘bakery’, ‘5’, ‘yes’*). Akin to a visual Turing test, answering any possible question about an image is one of the *‘holy grails’* of semantic understanding. VQA is directly applicable to a variety of applications of high societal impact that involve humans working in collaboration with machines to elicit and extract situationally-relevant information from visual data. Examples include aiding visually-impaired users in understanding their surroundings (*‘What temperature is my oven set to?’*), analysts in making decisions based on large quantities of surveillance data (*‘What kind of car did the man in the red shirt drive away in?’*), and users in interacting with a robot (*‘Is my laptop in my bedroom upstairs?’*). This research has the potential to fundamentally improve the way visually-impaired users live their daily lives, and revolutionize how society at large interacts with ever-growing visual data.

I provide below an overview of the **specific dimensions of VQA** that I study in this dissertation.

1.1 Free-form and Open-Ended VQA (chapter 3)

My colleagues and I introduced the task of free-form and open-ended VQA [27, 18]. In order to train and benchmark algorithms on the task of free-form and open-ended VQA, we collect and analyze a large scale dataset ($>0.25\text{M}$ images, $>0.76\text{M}$ questions, $\sim 10\text{M}$ answers) [27, 18]. The questions and answers in the dataset are provided by human workers on Amazon Mechanical Turk, on top of existing images [269, 516]. Unlike existing computer vision tasks which either represent single narrowly-defined problem (e.g., image classification, activity recognition), or are difficult to evaluate (e.g., image captioning), the questions in our VQA dataset require a potentially vast set of AI capabilities to answer (Fig. 3) – fine-grained recognition (e.g., *‘What kind of cheese is on the pizza?’*), object detection (e.g., *‘How many bikes are there?’*), and commonsense (e.g., *‘Does this person have 20/20 vision?’*). Moreover, VQA



Figure 1: Examples of free-form, open-ended questions in our VQA dataset.

is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We also develop and present experimental results of some baselines and methods for VQA. Finally, in order to push the state-of-the-art (SOTA) on VQA, we organize annual challenges and workshops on VQA and discuss the how these challenges and workshops improved the SOTA on VQA and benefitted the language and vision community in general.

1.2 Analyzing the Behavior of VQA Models (chapter 4)

After the release of our VQA dataset, a number of deep-learning models were proposed for VQA [27, 84, 495, 485, 208, 24, 465, 217, 280, 25, 400, 227, 150, 321, 197, 480, 483, 510, 381]. Curiously, the performance of most methods was clustered around 60-70% (compared to human performance of 83%) with a mere 5% gap between the top-9 entries on the VQA Challenge 2016. In order to identify the most fruitful directions for progress, we develop novel techniques for characterizing the behavior of VQA models [9]. We analyze several representative state-of-the-art VQA models [27, 280, 150], including the models developed by us [27] and present three novel findings that expand our understanding of VQA models – despite the progress, the VQA models are ‘*myopic*’ (tend to fail on sufficiently novel instances), often ‘*jump*



Figure 2: This figure illustrates outputs from a baseline model (SAN [495]) and the proposed model (GVQA [12]). For the given test questions, SAN predicts the prior answers from the training data for the respective question types, resulting in incorrect answers. However, GVQA, being more visually grounded than SAN, correctly answers the test questions.

to conclusions’ (converge on a predicted answer after ‘*listening*’ to just half the question), and are ‘*stubborn*’ (do not change their answers across images).

1.3 Overcoming Priors in VQA (chapter 5)

Motivated by the findings of our previous work [9] (and work by others [505, 168, 211]) that VQA models are heavily driven by superficial correlations in the training data and lack sufficient image grounding and compositionality, we address some of these issues by proposing:

a) **a new evaluation protocol (section 5.1).** We propose a new evaluation protocol for VQA – train and test sets have different prior distributions of answers for different question types (first few words of the question) [12]. Specifically, we create a new split of the VQA dataset [27] – Visual Question Answering under Changing Priors (VQA-CP). We evaluate several existing VQA models on the new split and find that their performance degrades significantly compared to the original VQA split. Thus, the proposed split can serve as a benchmark to evaluate the degree of visual groundedness in VQA models.

b) **a new model architecture (section 5.2).** We propose a novel Grounded

Visual Question Answering (GVQA) model [12] that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘*cheating*’ by primarily relying on priors in the training data. Specifically, GVQA explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question, enabling the model to more robustly generalize across different distributions of answers. GVQA significantly outperforms the baseline VQA model (SAN) [495] on VQA-CP. Fig. 2 illustrates outputs from GVQA and SAN. For the given test questions, SAN predicts the prior answers from the training data for the respective question types, resulting in incorrect answers. However, GVQA, being more visually grounded than SAN, correctly answers the test questions.

c) **a novel objective function (section 5.3)**. Although GVQA can be built on top of any existing VQA model, it does require non-trivial changes in the architecture. To address this issue, we propose a simple drop-in regularizer that can be added to any existing VQA model’s objective function [359]. To do this, we introduce a question-only model that takes the question encoding from the VQA model and must leverage language biases in order to succeed. We then pose training as an adversarial game between the VQA model and this question-only adversary – discouraging the VQA model from capturing language biases in its question encoding. This approach improves performance significantly for multiple base models (including GVQA), achieving state-of-the-art on VQA-CP.

1.4 Contributions

In this dissertation, we:

1. introduce the task of free-form and open-ended Visual Question Answering (VQA). We collect a large scale dataset (>0.25M images, >0.76M questions, ~10M answers) and make it publicly available (www.visualqa.org). We present

baselines and methods for VQA, and organize annual challenges and workshops to push the state-of-art on VQA.

2. develop novel techniques to characterize the behavior of VQA models. We analyze several representative VQA models and present three novel findings that expand our understanding of VQA models.
3. address the issue of VQA models being driven by superficial correlations in training data and lacking sufficient image grounding by proposing:
 - (a) a new evaluation protocol to evaluate the degree of visual groundedness in VQA models.
 - (b) a novel Grounded VQA (GVQA) model that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘cheating’ by primarily relying on priors in the training data.
 - (c) a novel adversarial regularization scheme that can be added to any existing VQA model’s objective function, without significantly changing the underlying VQA model’s architecture.

1.5 *List of Publications*

1. Sainandan Ramakrishnan, **A. Agrawal** and Stefan Lee. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. In *Neural Information Processing Systems (NIPS)*, 2018.
2. **A. Agrawal**, D. Batra, D. Parikh and A. Kembhavi. Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
3. G. Christie*, A. Laddha*, **A. Agrawal**, S. Antol, Y. Goyal, K. Kochersberger and D. Batra. Resolving Language and Vision Ambiguities Together: Joint

- Segmentation & Prepositional Attachment Resolution in Captioned Scenes. In the journal of *Computer Vision and Image Understanding (CVIU)*, 2017.
4. **A. Agrawal***, J. Lu*, S. Antol*, M. Mitchell, C. L. Zitnick, D. Parikh and D. Batra. VQA: Visual Question Answering. In Special Issue on *Combined Image and Language Understanding, International Journal of Computer Vision (IJCV)*, 2017.
 5. **A. Agrawal**, D. Batra and D. Parikh. Analyzing the Behavior of Visual Question Answering Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
 6. C. L. Zitnick, **A. Agrawal**, S. Antol, M. Mitchell, D. Batra and D. Parikh. Measuring Machine Intelligence Through Visual Question Answering. In *AI Magazine*, 2016.
 7. G. Christie*, A. Laddha*, **A. Agrawal**, S. Antol, Y. Goyal, K. Kochersberger and D. Batra. Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
 8. T. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, **A. Agrawal**, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C.L. Zitnick, D. Parikh, L. Vanderwende, M. Galley and M. Mitchell. Visual Storytelling. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.
 9. S. Antol*, **A. Agrawal***, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

CHAPTER II

RELATED WORK

In this chapter, I will discuss how our work on Visual Question Answering (VQA) is related to other research efforts in similar directions. I will first discuss related work on VQA, then Analyzing the Behavior of VQA Models followed by Overcoming Priors in VQA.

2.1 Visual Question Answering (VQA)

VQA Efforts. Several recent papers have studied visual question answering [158, 286, 439, 51]. However, unlike our work, these are fairly restricted (sometimes synthetic) settings with small datasets. For instance, [286] only considers questions whose answers come from a predefined closed world of 16 basic colors or 894 object categories. [158] also considers questions generated from templates from a fixed vocabulary of objects, attributes, relationships between objects, *etc.* In contrast, our proposed task involves *open-ended, free-form* questions and answers provided by humans. Our goal is to increase the diversity of knowledge and kinds of reasoning needed to provide correct answers. Critical to achieving success on this more difficult and unconstrained task, our VQA dataset is *two orders of magnitude* larger than [158, 286] (>250,000 *vs.* 2,591 and 1,449 images respectively). The proposed VQA task has connections to other related work: [439] has studied joint parsing of videos and corresponding text to answer queries on two datasets containing 15 video clips each. [51] uses crowdsourced workers to answer questions about visual content asked by visually-impaired users. In concurrent work, [289] proposed combining an LSTM for the question with a CNN for the image to generate an answer. In their model, the LSTM question representation is conditioned on the CNN image features

at each time step, and the final LSTM hidden state is used to sequentially decode the answer phrase. In contrast, the model developed by us explores “late fusion” – *i.e.*, the LSTM question representation and the CNN image features are computed independently, *fused* via an element-wise multiplication, and then passed through fully-connected layers to generate a softmax distribution over output answer classes. [272] generates abstract scenes to capture visual common sense relevant to answering (purely textual) fill-in-the-blank and visual paraphrasing questions. [378] and [448] use visual information to assess the plausibility of common sense assertions. [499] introduced a dataset of 10k images and prompted captions that describe specific aspects of a scene (*e.g.*, individual objects, what will happen next). Concurrent with our work, [156] collected questions & answers in Chinese (later translated to English by humans) for COCO images. [369] automatically generated four types of questions (object, count, color, location) using COCO captions.

Text-based Q&A is a well studied problem in the NLP and text processing communities (recent examples being [129, 128, 473, 372]). Other related textual tasks include sentence completion (*e.g.*, [372] with multiple-choice answers). These approaches provide inspiration for VQA techniques. One key concern in text is the *grounding* of questions. For instance, [473] synthesized textual descriptions and QA-pairs grounded in a simulation of actors and objects in a fixed set of locations. VQA is naturally grounded in images – requiring the understanding of both text (questions) and vision (images). Our questions are generated by humans, making the need for commonsense knowledge and complex reasoning more essential.

Describing Visual Content. Related to VQA are the tasks of image tagging [112, 243], image captioning [247, 134, 309, 89, 131, 451, 116, 222, 290, 229] and video captioning [374, 172], where words or sentences are generated to describe visual content. While these tasks require both visual and semantic knowledge, captions can often be non-specific (*e.g.*, observed by [451]). The questions in VQA require detailed

specific information about the image for which generic image captions are of little use [51].

Other Vision+Language Tasks. Several recent papers have explored tasks at the intersection of vision and language that are easier to evaluate than image captioning, such as coreference resolution [235, 361] or generating referring expressions [225, 312] for a particular object in an image that would allow a human to identify which object is being referred to (*e.g.*, “the one in a red shirt”, “the dog on the left”). While task-driven and concrete, a limited set of visual concepts (*e.g.*, color, location) tend to be captured by referring expressions. As we demonstrate, a richer variety of visual concepts emerge from visual questions and their answers.

2.2 *Analyzing the Behavior of VQA Models*

Our work is inspired by previous works that diagnose the failure modes of models for different tasks. [223] constructed a series of oracles to measure the performance of a character level language model. [189] provided analysis tools to facilitate detailed and meaningful investigation of object detector performance. Our work aims to perform behavior analyses as a first step towards diagnosing errors for VQA.

[495] categorize the errors made by their VQA model into four categories – model focuses attention on incorrect regions, model focuses attention on appropriate regions but predicts incorrect answers, predicted answers are different from labels but might be acceptable, labels are wrong. While these are coarse but useful failure modes, we are interested in understanding the behavior of VQA models along specific dimensions – whether they generalize to novel instances, whether they listen to the entire question, whether they look at the image.

2.3 *Overcoming Priors in VQA*

Countering Priors in VQA: In order to counter the language priors in the VQA v1 dataset, [168] balance every question by collecting complementary images for every question. Thus, for every question in the proposed VQA v2 dataset, there are two similar images with different answers to the question. By construction, language priors are significantly weaker in the VQA v2 dataset. However, the train and test distributions are still similar, unlike our work where the train and test answer distributions are by design different (section 5.1). So, leveraging priors from the train set will still benefit the model at test time. [505] balance the yes/no questions on abstract scenes from the VQA v1 dataset in a similar manner. More recently, [216] propose two new evaluation metrics that compensate for the skewed distribution of question types and for the skewed distribution of answers within each question type in the test set. As a remedy for machines using “shortcuts” to solve multiple-choice VQA, [79] describe several principles for automatic construction of good decoys (the incorrect candidate answers). [80] study cross-dataset adaptation for VQA. They propose an algorithm for adapting a VQA model trained on one dataset to apply to another dataset with different statistical distribution. All these works indicate that there is an increasing interest in the community to focus on models that are less driven by training priors and are more visually grounded.

Compositionality. Related to the ability to generalize across different answer distributions is the ability to generalize to novel compositions of known concepts learned during training. Compositionality has been studied in various forms in the vision community. Zero-shot object recognition using attributes is based on the idea of composing attributes to detect novel object categories [256, 206]. [31] have studied compositionality in the domain of image captioning by focusing on structured representations (subject-relation-object triplets). We study compositionality for visual question answering where the questions and answers are open-ended and in free-form

natural language. The work closest to us is [211] where they study compositionality in the domain of VQA. However, their dataset (images as well as questions) is synthetic and has only limited number of objects and attributes. On the contrary, our C-VQA splits consist of real images and questions (asked by humans) and hence involve a variety of objects and attributes, as well as activities, scenes, etc. Andreas et al. [24, 25] have developed compositional models for VQA that consist of different modules each specialized for a particular task. These modules can be composed together based on the question structure to create a model architecture for the given question. Although, compositional by design, these models have not been evaluated specifically for compositionality. Our C-VQA splits can be used to evaluate such models to test the degree of compositionality. In fact, we report the performance of Neural Module Networks on our C-VQA splits and find that its performance degrades significantly from the original VQA setting to the proposed C-VQA setting (section ??).

Zero-shot VQA has also been explored in [428]. They study a setting for VQA where the test questions (the question string itself or the multiple choices) contain at least one unseen word. [360] propose answering questions about unknown objects (*e.g.*, ‘*Is the dog black and white?*’ where ‘*dog*’ is never seen in training questions or answers). These are orthogonal efforts to our work in that our focus is not in studying if unseen words/concepts can be recognized during testing. We are instead interested in studying – 1) the extent to which a model is visually grounded by evaluating its ability to generalize to a different answer distribution for each question type, 2) the extent to which a model is able to answer questions about unseen compositions of seen concepts. In both the splits proposed by us (VQA-CP and C-VQA), we ensure that concepts seen during test time are present during training to the extent possible.

Adversarial Learning. Generative Adversarial Networks (GANs) [164] have received significant recent interest for their ability to model complex distributions – finding use in a variety of image and language generation tasks [164, 356, 504, 102,

307]. Recently, other adversarial training schemes have been proposed to encourage various forms of invariance in intermediate model representations [257, 277, 442].

Most related to our work on adversarial regularization for VQA (section 5.3), Lample *et al.* [257] introduce an autoencoder framework with an adversarial loss for attribute-based image manipulation. Given an input image and a set of attributes (*e.g.* a photo of a person and their gender or age), the task is to manipulate the image such that it has the desired attributes. Unfortunately, without multiple pairings of the same image with different attributes, it is challenging to learn disentangled image representations that generalize to new input-attribute combinations. An adversarial model is introduced that is trained to predict attributes from the input image encoding alone. In combating this adversary, the image encoder model learns to produce attribute invariant image encodings. This improves generalization by forcing the attribute-augmented decoder to meaningfully rely on input attributes to accurately reproduce input images.

Similarly, the question-only adversary in our work (section 5.3), encourages the VQA question encoder to remove answer-discriminative features from the question representation. However, breaking the parallels with [257], the answer themselves are not added back as inputs to controllably recondition the model on these features. Rather, the VQA model must rely on the combination of question and image features to recover the answer information. In this way, the language-level answer information (*e.g.* that most grass is green) is removed from the question and instance-specific information from the image must be used instead. We take this notion further by leveraging the question-only adversary to estimate and directly maximize the change in confidence after observing the image, which we show provides substantial benefits when paired with the question-only adversary.

CHAPTER III

VISUAL QUESTION ANSWERING (VQA)

3.1 *Introduction*

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [131, 89, 116, 290, 229, 222, 451]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word n -gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as “AI-complete” as desired.

What makes for a compelling “AI-complete” task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a single sub-domain (such as CV) and (ii) have a well-defined *quantitative evaluation metric* to track progress. For some tasks, such as image captioning, automatic evaluation is still a difficult and open research problem [447, 125, 186].

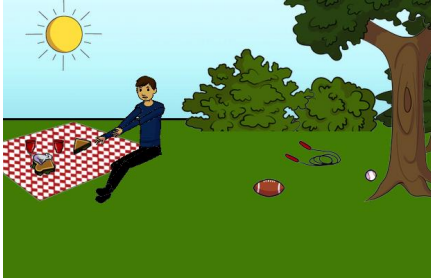
In this chapter, we introduce the task of *free-form* and *open-ended* Visual Question Answering (VQA). A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output. This goal-driven task is applicable to scenarios encountered when visually-impaired users [51] or intelligence analysts actively elicit visual information. Example questions are shown in Fig. 3.



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Figure 3: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Open-ended questions require a potentially vast set of AI capabilities to answer – fine-grained recognition (*e.g.*, “What kind of cheese is on the pizza?”), object detection (*e.g.*, “How many bikes are there?”), activity recognition (*e.g.*, “Is this man crying?”), knowledge base reasoning (*e.g.*, “Is this a vegetarian pizza?”), and commonsense reasoning (*e.g.*, “Does this person have 20/20 vision?”, “Is this person expecting company?”). VQA [158, 286, 439, 51] is also amenable to automatic quantitative evaluation, making it possible to effectively track progress on this task. While the answer to many questions is simply “yes” or “no”, the process for determining a correct answer is typically far from trivial (*e.g.* in Fig. 3, “Does this person have 20/20 vision?”). Moreover, since questions about images often tend to seek specific information, simple one-to-three word answers are sufficient for many questions. In such scenarios, we can easily evaluate a proposed algorithm by the number of questions it answers correctly. In this work, we present both an open-ended answering

task and a multiple-choice task [372, 271]. Unlike the open-ended task that requires a free-form response, the multiple-choice task only requires an algorithm to pick from a predefined list of possible answers.

We present a large dataset that contains 204,721 images from the MS COCO dataset [269] and a newly created abstract scene dataset [516, 29] that contains 50,000 scenes. The MS COCO dataset has images depicting diverse and complex scenes that are effective at eliciting compelling and diverse questions. We collected a new dataset of “realistic” abstract scenes to enable research focused only on the high-level reasoning required for VQA by removing the need to parse real images. Three questions were collected for each image or scene. Each question was answered by ten subjects along with their confidence. The dataset contains over 760K questions with around 10M answers.

While the use of open-ended questions offers many benefits, it is still useful to understand the types of questions that are being asked and which types various algorithms may be good at answering. To this end, we analyze the types of questions asked and the types of answers provided. Through several visualizations, we demonstrate the astonishing diversity of the questions asked. We also explore how the information content of questions and their answers differs from image captions. For baselines, we offer several approaches that use a combination of both text and state-of-the-art visual features [243]. As part of the VQA initiative, we have been organizing annual challenges and associated workshops to discuss state-of-the-art methods and best practices.

VQA poses a rich set of challenges, many of which have been viewed as the holy grail of automatic image understanding and AI in general. However, it includes as building blocks several components that the CV, NLP, and KR [72, 88, 261, 273, 57] communities have made significant progress on during the past few decades. VQA provides an attractive balance between pushing the state of the art, while being

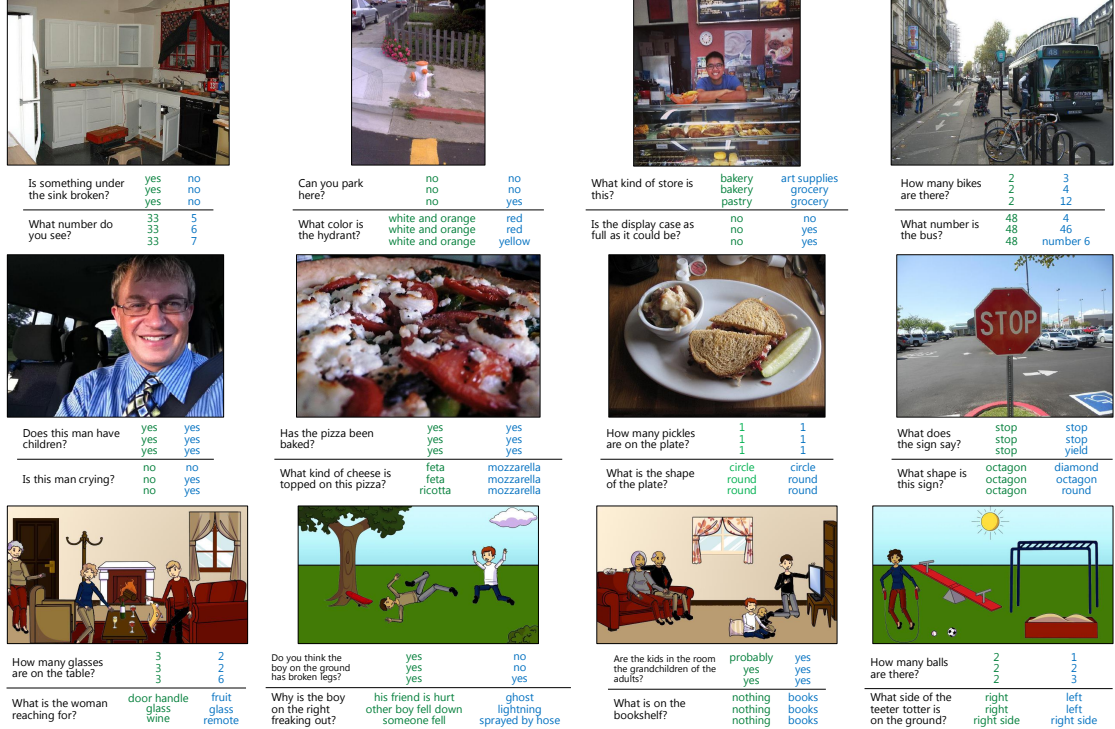


Figure 4: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

accessible enough for the communities to start making progress on the task.

3.2 VQA Dataset Collection

We now describe the Visual Question Answering (VQA) dataset. We begin by describing the real images and abstract scenes used to collect the questions. Next, we describe our process of collecting questions and their corresponding answers. Analysis of the questions and answers gathered as well as baselines' & methods' results are provided in following sections.

Real Images. We use the 123,287 training and validation images and 81,434 test images from the Microsoft Common Objects in Context (MS COCO) [269] dataset. The MS COCO dataset was gathered to find images containing multiple objects and rich contextual information. Given the visual complexity of these images, they

are well-suited for our VQA task. The more diverse our collection of images, the more diverse, comprehensive, and interesting the resultant set of questions and their answers.

Abstract Scenes. The VQA task with real images requires the use of complex and often noisy visual recognizers. To attract researchers interested in exploring the high-level reasoning required for VQA, but not the low-level vision tasks, we create a new abstract scenes dataset [29, 516, 517, 518] containing 50K scenes. The dataset contains 20 “paperdoll” human models [29] spanning genders, races, and ages with 8 different expressions. The limbs are adjustable to allow for continuous pose variations. The clipart may be used to depict both indoor and outdoor scenes. The set contains over 100 objects and 31 animals in various poses. The use of this clipart enables the creation of more realistic scenes (see bottom row of Fig. 4) that more closely mirror real images than previous papers [516, 517, 518]. See the appendix for the user interface, additional details, and examples.

Splits. For real images, we follow the same train/val/test split strategy as the MC COCO dataset [269] (including test-dev, test-standard, test-challenge, test-reserve). For the VQA challenge (see section ??), test-dev is used for debugging and validation experiments and allows for unlimited submission to the evaluation server. Test-standard is the ‘default’ test data for the VQA competition. When comparing to the state of the art (e.g., in papers), results should be reported on test-standard. Test-standard is also used to maintain a public leaderboard that is updated upon submission. Test-reserve is used to protect against possible overfitting. If there are substantial differences between a method’s scores on test-standard and test-reserve, this raises a red-flag and prompts further investigation. Results on test-reserve are not publicly revealed. Finally, test-challenge is used to determine the winners of the challenge.

For abstract scenes, we created splits for standardization, separating the scenes

into 20K/10K/20K for train/val/test splits, respectively. There are no subsplits (test-dev, test-standard, test-challenge, test-reserve) for abstract scenes.

Captions. The MS COCO dataset [269, 86] already contains five single-sentence captions for all images. We also collected five single-captions for all abstract scenes using the same user interface¹ for collection.

Questions. Collecting interesting, diverse, and well-posed questions is a significant challenge. Many simple questions may only require low-level computer vision knowledge, such as “What color is the cat?” or “How many chairs are present in the scene?”. However, we also want questions that require commonsense knowledge about the scene, such as “What sound does the pictured animal make?”. Importantly, questions should also *require* the image to correctly answer and not be answerable using just commonsense information, *e.g.*, in Fig. 3, “What is the mustache made of?”. By having a wide variety of question types and difficulty, we may be able to measure the continual progress of both visual understanding and commonsense reasoning.

We tested and evaluated a number of user interfaces for collecting such “interesting” questions. Specifically, we ran pilot studies asking human subjects to ask questions about a given image that they believe a “toddler”, “alien”, or “smart robot” would have trouble answering. We found the “smart robot” interface to elicit the most interesting and diverse questions. As shown in the appendix, our final interface stated:

¹<https://github.com/tylin/coco-ui>

“We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people’s expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot!

Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image.”

To bias against generic image-independent questions, subjects were instructed to ask questions that *require* the image to answer.

The same user interface was used for both the real images and abstract scenes. In total, three questions from unique workers were gathered for each image/scene. When writing a question, the subjects were shown the previous questions already asked for that image to increase the question diversity. In total, the dataset contains over $\sim 0.76\text{M}$ questions.

Answers. Open-ended questions result in a diverse set of possible answers. For many questions, a simple “yes” or “no” response is sufficient. However, other questions may require a short phrase. Multiple different answers may also be correct. For instance, the answers “white”, “tan”, or “off-white” may all be correct answers to the same question. Human subjects may also disagree on the “correct” answer, *e.g.*, some saying “yes” while others say “no”. To handle these discrepancies, we gather *10 answers for each question from unique workers*, while also ensuring that the worker answering a question did not ask it. We ask the subjects to provide answers that are “a brief phrase and not a complete sentence. Respond matter-of-factly and avoid using conversational language or inserting your opinion.” In addition to answering the questions, the subjects were asked “Do you think you were able to answer the question correctly?” and given the choices of “no”, “maybe”, and “yes”. See the appendix for more details about the user interface to collect answers. See Section 3.3

for an analysis of the answers provided.

For testing, we offer two modalities for answering the questions: (i) **open-ended** and (ii) **multiple-choice**.

For the open-ended task, the generated answers are evaluated using the following accuracy metric:

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

i.e., an answer is deemed 100% accurate if at least 3 workers provided that exact answer.² Before comparison, all responses are made lowercase, numbers converted to digits, and punctuation & articles removed. We avoid using soft metrics such as Word2Vec [303], since they often group together words that we wish to distinguish, such as “left” and “right”. We also avoid using evaluation metrics from machine translation such as BLEU and ROUGE because such metrics are typically applicable and reliable for sentences containing multiple words. In VQA, most answers (89.32%) are single word; thus there no high-order n-gram matches between predicted answers and ground-truth answers, and low-order n-gram matches degenerate to exact-string matching. Moreover, these automatic metrics such as BLEU and ROUGE have been found to poorly correlate with human judgement for tasks such as image caption evaluation [87].

For multiple-choice task, 18 candidate answers are created for each question. As with the open-ended task, the accuracy of a chosen option is computed based on the number of human subjects who provided that answer (divided by 3 and clipped at 1). We generate a candidate set of correct and incorrect answers from four sets of answers: **Correct:** The most common (out of ten) correct answer. **Plausible:** To generate incorrect, but still plausible answers we ask three subjects to answer the questions without seeing the image. See the appendix for more details about

²In order to be consistent with ‘human accuracies’ reported in Section ??, machine accuracies are averaged over all $\binom{10}{9}$ sets of human annotators

the user interface to collect these answers. If three unique answers are not found, we gather additional answers from nearest neighbor questions using a bag-of-words model. The use of these answers helps ensure the image, and not just commonsense knowledge, is necessary to answer the question. **Popular:** These are the 10 most popular answers. For instance, these are “yes”, “no”, “2”, “1”, “white”, “3”, “red”, “blue”, “4”, “green” for real images. The inclusion of the most popular answers makes it more difficult for algorithms to infer the type of question from the set of answers provided, *i.e.*, learning that it is a “yes or no” question just because “yes” and “no” are present in the answers. **Random:** Correct answers from random questions in the dataset. To generate a total of 18 candidate answers, we first find the union of the correct, plausible, and popular answers. We include random answers until 18 unique answers are found. The order of the answers is randomized. Example multiple choice questions are in the appendix.

Note that all 18 candidate answers are unique. But since 10 different subjects answered every question, it is possible that more than one of those 10 answers be present in the 18 choices. In such cases, according to the accuracy metric, multiple options could have a non-zero accuracy.

3.3 VQA Dataset Analysis

In this section, we provide an analysis of the questions and answers in the VQA train dataset. To gain an understanding of the types of questions asked and answers provided, we visualize the distribution of question types and answers. We also explore how often the questions may be answered without the image using just commonsense information. Finally, we analyze whether the information contained in an image caption is sufficient to answer the questions.

The dataset includes 614,163 questions and 7,984,119 answers (including answers provided by workers with and without looking at the image) for 204,721 images from

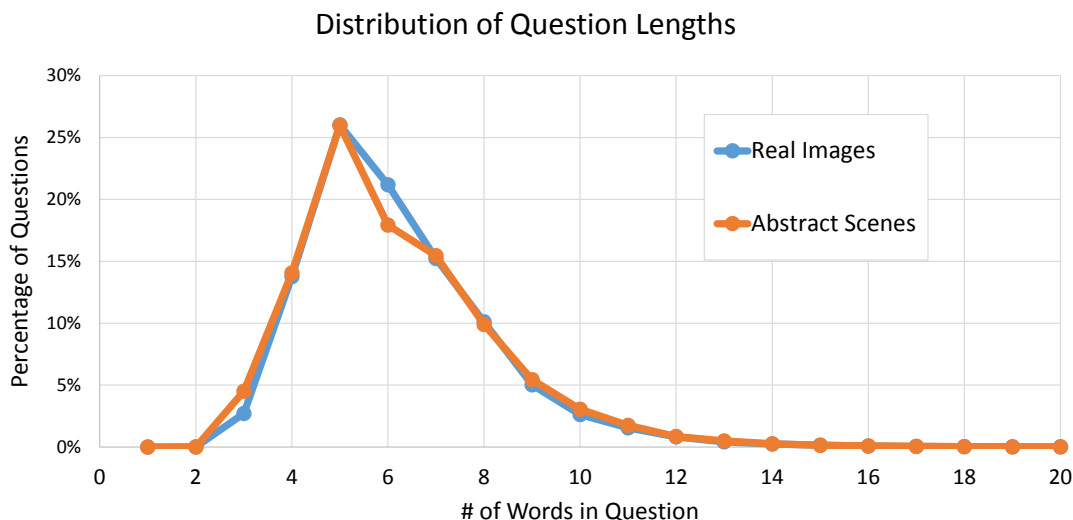


Figure 6: Percentage of questions with different word lengths for real images and abstract scenes.

of possible answers. See the appendix for visualizations for “What is...” questions.

Lengths. Fig. 6 shows the distribution of question lengths. We see that most questions range from four to ten words.

3.3.2 Answers

Typical Answers. Fig. 7 (top) shows the distribution of answers for several question types. We can see that a number of question types, such as “Is the...”, “Are...”, and “Does...” are typically answered using “yes” and “no” as answers. Other questions such as “What is...” and “What type...” have a rich diversity of responses. Other question types such as “What color...” or “Which...” have more specialized responses, such as colors, or “left” and “right”. See the appendix for a list of the most popular answers.

Lengths. Most answers consist of a single word, with the distribution of answers containing one, two, or three words, respectively being 89.32%, 6.91%, and 2.74% for real images and 90.51%, 5.89%, and 2.49% for abstract scenes. The brevity of answers is not surprising, since the questions tend to elicit specific information from the images. This is in contrast with image captions that generically describe

[illegible]

25

the entire image and hence tend to be longer. The brevity of our answers makes automatic evaluation feasible. While it may be tempting to believe the brevity of the answers makes the problem easier, recall that they are human-provided open-ended answers to open-ended questions. The questions typically require complex reasoning to arrive at these deceptively simple answers (see Fig. 4). There are currently 23,234 unique one-word answers in our dataset for real images and 3,770 for abstract scenes.

‘Yes/No’ and ‘Number’ Answers. Many questions are answered using either “yes” or “no” (or sometimes “maybe”) – 38.37% and 40.66% of the questions on real images and abstract scenes respectively. Among these ‘yes/no’ questions, there is a bias towards “yes” – 58.83% and 55.86% of ‘yes/no’ answers are “yes” for real images and abstract scenes. Question types such as “How many...” are answered using numbers – 12.31% and 14.48% of the questions on real images and abstract scenes are ‘number’ questions. “2” is the most popular answer among the ‘number’ questions, making up 26.04% of the ‘number’ answers for real images and 39.85% for abstract scenes.

Subject Confidence. When the subjects answered the questions, we asked “Do you think you were able to answer the question correctly?”. Fig. 8 shows the distribution of responses. A majority of the answers were labeled as confident for both real images and abstract scenes.

Inter-human Agreement. Does the self-judgment of confidence correspond to the answer agreement between subjects? Fig. 8 shows the percentage of questions in which (i) 7 or more, (ii) 3 – 7, or (iii) less than 3 subjects agree on the answers given their average confidence score (0 = not confident, 1 = confident). As expected, the agreement between subjects increases with confidence. However, even if all of the subjects are confident the answers may still vary. This is not surprising since some answers may vary, yet have very similar meaning, such as “happy” and “joyful”.

As shown in Table 3.3.3 (Question + Image), there is significant inter-human

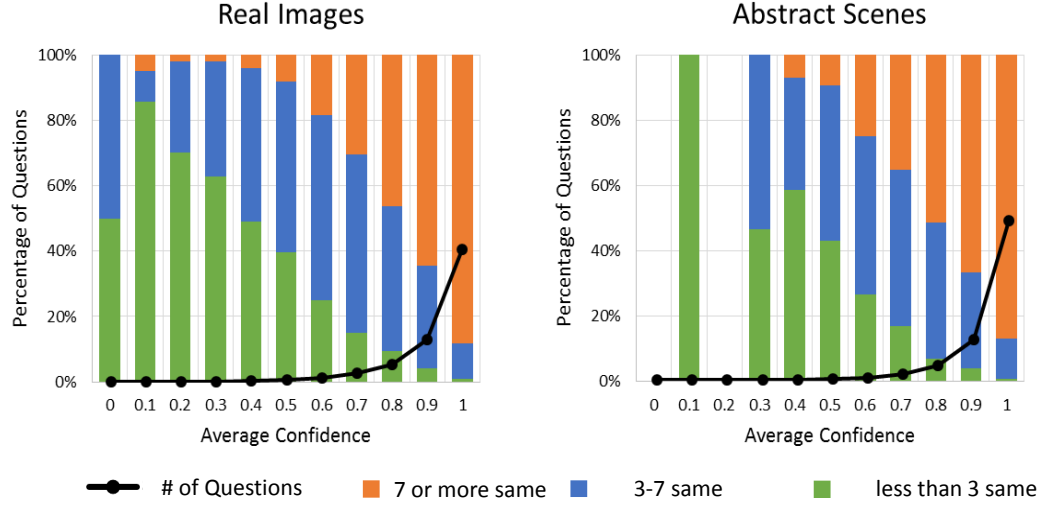


Figure 8: Number of questions per average confidence score (0 = not confident, 1 = confident) for real images and abstract scenes (black lines). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same (color bars).

| | | | | |
|--|---|---|--|---|
| 3-4 (15.3%) Is that a bird in the sky? What color is the shoe? How many zebras are there? Is there food on the table? Is this man wearing shoes? | 5-8 (39.7%) How many pizzas are shown? What are the sheep eating? What color is his hair? What sport is being played? Name one ingredient in the skillet. | 9-12 (28.4%) Where was this picture taken? What ceremony does the cake commemorate? Are these boats too tall to fit under the bridge? What is the name of the white shape under the batter? Is this at the stadium? | 13-17 (11.2%) Is he likely to get mugged if he walked down a dark alleyway like this? Is this a vegetarian meal? What type of beverage is in the glass? Can you name the performer in the purple costume? Besides these humans, what other animals eat here? | 18+ (5.5%) What type of architecture is this? Is this a Flemish bricklaying pattern? How many calories are in this pizza? What government document is needed to partake in this activity? What is the make and model of this vehicle? |
|--|---|---|--|---|

Figure 9: Example questions judged by Mturk workers to be answerable by different age groups. The percentage of questions falling into each age group is shown in parentheses.

agreement in the answers for both real images (83.30%) and abstract scenes (87.49%). Note that on average each question has 2.70 unique answers for real images and 2.39 for abstract scenes. The agreement is significantly higher ($> 95\%$) for “yes/no” questions and lower for other questions ($< 76\%$), possibly due to the fact that we perform exact string matching and do not account for synonyms, plurality, *etc.* Note that the automatic determination of synonyms is a difficult problem, since the level of answer granularity can vary across questions.

3.3.3 Commonsense Knowledge

Is the Image Necessary? Clearly, some questions can sometimes be answered correctly using commonsense knowledge alone without the need for an image, *e.g.*, “What is the color of the fire hydrant?”. We explore this issue by asking three subjects to answer the questions *without seeing the image* (see the examples in blue in Fig. 4). In Table 3.3.3 (Question), we show the percentage of questions for which the correct answer is provided over all questions, “yes/no” questions, and the other questions that are not “yes/no”. For “yes/no” questions, the human subjects respond better than chance. For other questions, humans are only correct about 21% of the time. This demonstrates that understanding the visual information is critical to VQA and that commonsense information alone is not sufficient.

To show the qualitative difference in answers provided with and without images, we show the distribution of answers for various question types in Fig. 7 (bottom). The distribution of colors, numbers, and even “yes/no” responses is surprisingly different for answers with and without images.

Which Questions Require Common Sense? In order to identify questions that require commonsense reasoning to answer, we conducted two AMT studies (on a subset 10K questions from the real images of VQA trainval) asking subjects –

1. Whether or not they believed a question required commonsense to answer the question, and
2. The youngest age group that they believe a person must be in order to be able to correctly answer the question – toddler (3-4), younger child (5-8), older child (9-12), teenager (13-17), adult (18+).

Each question was shown to 10 subjects. We found that for 47.43% of questions 3 or more subjects voted ‘yes’ to commonsense, (18.14%: 6 or more). In the ‘perceived human age required to answer question’ study, we found the following distribution of responses: toddler: 15.3%, younger child: 39.7%, older child: 28.4%, teenager: 11.2%,

adult: 5.5%. In Figure 9 we show several questions for which a majority of subjects picked the specified age range. Surprisingly the perceived age needed to answer the questions is fairly well distributed across the different age ranges. As expected the questions that were judged answerable by an adult (18+) generally need specialized knowledge, whereas those answerable by a toddler (3-4) are more generic.

We measure the degree of commonsense required to answer a question as the percentage of subjects (out of 10) who voted “yes” in our “whether or not a question requires commonsense” study. A fine-grained breakdown of average age and average degree of common sense (on a scale of 0 – 100) required to answer a question is shown in Table ?? . The average age and the average degree of commonsense across all questions is 8.92 and 31.01% respectively.

It is important to distinguish between:

1. How old someone needs to be to be able to answer a question correctly, and
2. How old people *think* someone needs to be to be able to answer a question correctly.

Our age annotations capture the latter – perceptions of MTurk workers in an uncontrolled environment. As such, the relative ordering of question types in Table ?? is more important than absolute age numbers. The two rankings of questions in terms of common sense required according to the two studies were largely correlated (Pearson’s rank correlation: 0.58).

3.3.4 Captions *vs.* Questions

Do generic image captions provide enough information to answer the questions? Table 3.3.3 (Question + Caption) shows the percentage of questions answered correctly when human subjects are given the question and a human-provided caption describing the image, but not the image. As expected, the results are better than when humans are shown the questions alone. However, the accuracies are significantly

Table 1: Test-standard accuracy of human subjects when asked to answer the question without seeing the image (Question), seeing just a caption of the image and not the image itself (Question + Caption), and seeing the image (Question + Image). Results are shown for all questions, “yes/no” & “number” questions, and other questions that are neither answered “yes/no” nor number. All answers are free-form and not multiple-choice. *These accuracies are evaluated on a subset of 3K train questions (1K images).

| Dataset | Input | All | Yes/No | Number | Other |
|----------|---------------------|-------|--------|--------|-------|
| Real | Question | 40.81 | 67.60 | 25.77 | 21.22 |
| | Question + Caption* | 57.47 | 78.97 | 39.68 | 44.41 |
| | Question + Image | 83.30 | 95.77 | 83.39 | 72.67 |
| Abstract | Question | 43.27 | 66.65 | 28.52 | 23.66 |
| | Question + Caption* | 54.34 | 74.70 | 41.19 | 40.18 |
| | Question + Image | 87.49 | 95.96 | 95.04 | 75.33 |

lower than when subjects are shown the actual image. This demonstrates that in order to answer the questions correctly, deeper image understanding (beyond what image captions typically capture) is necessary. In fact, we find that the distributions of nouns, verbs, and adjectives mentioned in captions is statistically significantly different from those mentioned in our questions + answers (Kolmogorov-Smirnov test, $p < .001$) for both real images and abstract scenes. See the appendix for details.

3.4 VQA Baselines and Methods

In this section, we explore the difficulty of the VQA dataset for the MS COCO images using several baselines and novel methods. We train on VQA train+val. Unless stated otherwise, all human accuracies are on test-standard, machine accuracies are on test-dev, and results involving human captions (in gray font) are trained on train and tested on val (because captions are not available for test).

3.4.1 Baselines

We implemented the following baselines:

1. **random:** We randomly choose an answer from the top 1K answers of the VQA

train/val dataset.

2. **prior (“yes”):** We always select the most popular answer (“yes”) for both the open-ended and multiple-choice tasks. Note that “yes” is always one of the choices for the multiple-choice questions.
3. **per Q-type prior:** For the open-ended task, we pick the most popular answer per question type (see the appendix for details). For the multiple-choice task, we pick the answer (from the provided choices) that is most similar to the picked answer for the open-ended task using cosine similarity in Word2Vec[303] feature space.
4. **nearest neighbor:** Given a test image, question pair, we first find the K nearest neighbor questions and associated images from the training set. See appendix for details on how neighbors are found. Next, for the open-ended task, we pick the most frequent ground truth answer from this set of nearest neighbor question, image pairs. Similar to the “per Q-type prior” baseline, for the multiple-choice task, we pick the answer (from the provided choices) that is most similar to the picked answer for the open-ended task using cosine similarity in Word2Vec[303] feature space.

3.4.2 Methods

For our methods, we develop a 2-channel vision (image) + language (question) model that culminates with a softmax over K possible outputs. We choose the top $K = 1000$ most frequent answers as possible outputs. This set of answers covers 82.67% of the train+val answers. We describe the different components of our model below:

Image Channel: This channel provides an embedding for the image. We experiment with two embeddings –

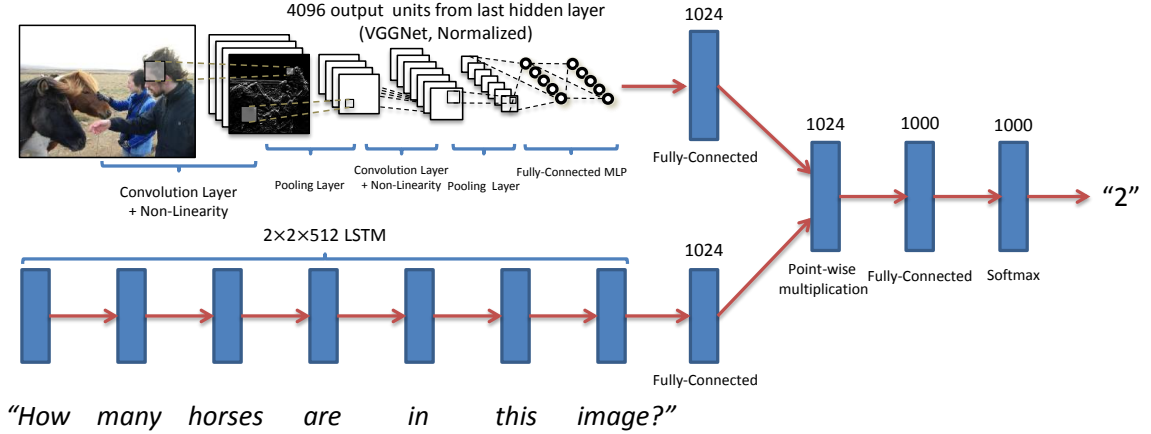


Figure 10: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [408] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

1. **I:** The activations from the last hidden layer of VGGNet [408] are used as 4096-dim image embedding.
2. **norm I:** These are ℓ_2 normalized activations from the last hidden layer of VGGNet [408].

Question Channel: This channel provides an embedding for the question. We experiment with three embeddings –

1. **Bag-of-Words Question (BoW Q):** The top 1,000 words in the questions are used to create a bag-of-words representation. Since there is a strong correlation between the words that start a question and the answer (see Fig. 7), we find the top 10 first, second, and third words of the questions and create a 30 dimensional bag-of-words representation. These features are concatenated to get a 1,030-dim embedding for the question.
2. **LSTM Q:** An LSTM with one hidden layer is used to obtain 1024-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of last cell state and last hidden state representations (each being

512-dim) from the hidden layer of the LSTM. Each question word is encoded with 300-dim embedding by a fully-connected layer + tanh non-linearity which is then fed to the LSTM. The input vocabulary to the embedding layer consists of all the question words seen in the training dataset.

3. **deeper LSTM Q**: An LSTM with two hidden layers is used to obtain 2048-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of last cell state and last hidden state representations (each being 512-dim) from each of the two hidden layers of the LSTM. Hence 2 (hidden layers) \times 2 (cell state and hidden state) \times 512 (dimensionality of each of the cell states, as well as hidden states) in Fig. 10. This is followed by a fully-connected layer + tanh non-linearity to transform 2048-dim embedding to 1024-dim. The question words are encoded in the same way as in LSTM Q.

Multi-Layer Perceptron (MLP): The image and question embeddings are combined to obtain a single embedding.

1. For **BoW Q + I** method, we simply concatenate the BoW Q and I embeddings.
2. For **LSTM Q + I**, and **deeper LSTM Q + norm I** (Fig. 10) methods, the image embedding is first transformed to 1024-dim by a fully-connected layer + tanh non-linearity to match the LSTM embedding of the question. The transformed image and LSTM embeddings (being in a common space) are then fused via element-wise multiplication.

This combined image + question embedding is then passed to an MLP – a fully connected neural network classifier with 2 hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh non-linearity, followed by a softmax layer to obtain a distribution over K answers. The entire model is learned end-to-end with a cross-entropy loss. VGGNet parameters are frozen to those learned for ImageNet classification and not fine-tuned in the image channel.

We also experimented with providing captions as input to our model. Similar to

Table 2: Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val). See text for details.

| | Open-Ended | | | | Multiple-Choice | | | |
|------------------------|--------------|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|
| | All | Yes/No | Number | Other | All | Yes/No | Number | Other |
| prior (“yes”) | 29.66 | 70.81 | 00.39 | 01.15 | 29.66 | 70.81 | 00.39 | 01.15 |
| per Q-type prior | 37.54 | 71.03 | 35.77 | 09.38 | 39.45 | 71.02 | 35.86 | 13.34 |
| nearest neighbor | 42.70 | 71.89 | 24.36 | 21.94 | 48.49 | 71.94 | 26.00 | 33.56 |
| BoW Q | 48.09 | 75.66 | 36.70 | 27.14 | 53.68 | 75.71 | 37.05 | 38.64 |
| I | 28.13 | 64.01 | 00.42 | 03.77 | 30.53 | 69.87 | 00.45 | 03.76 |
| BoW Q + I | 52.64 | 75.55 | 33.67 | 37.37 | 58.97 | 75.59 | 34.35 | 50.33 |
| LSTM Q | 48.76 | 78.20 | 35.68 | 26.59 | 54.75 | 78.22 | 36.82 | 38.78 |
| LSTM Q + I | 53.74 | 78.94 | 35.24 | 36.42 | 57.17 | 78.95 | 35.80 | 43.41 |
| deeper LSTM Q | 50.39 | 78.41 | 34.68 | 30.03 | 55.88 | 78.45 | 35.91 | 41.13 |
| deeper LSTM Q + norm I | 57.75 | 80.50 | 36.77 | 43.08 | 62.70 | 80.52 | 38.22 | 53.01 |
| Caption | 26.70 | 65.50 | 02.03 | 03.86 | 28.29 | 69.79 | 02.06 | 03.82 |
| BoW Q + C | 54.70 | 75.82 | 40.12 | 42.56 | 59.85 | 75.89 | 41.16 | 52.53 |

Table 3.3.3, we assume that a human-generated caption is given as input. We use a bag-of-words representation containing the 1,000 most popular words in the captions as the caption embedding (**Caption**). For **BoW Question + Caption (BoW Q + C)** method, we simply concatenate the BoW Q and C embeddings.

For testing, we report the result on two different tasks: open-ended selects the answer with highest activation from all possible K answers and multiple-choice picks the answer that has the highest activation from the potential answers.

3.4.3 Results

Table 2 shows the accuracy of our baselines and methods for both the open-ended and multiple-choice tasks on the VQA test-dev for real images.

As expected, the vision-alone model (I) that completely ignores the question performs rather poorly (open-ended: 28.13% / multiple-choice: 30.53%). In fact, on open-ended task, the vision-alone model (I) performs worse than the prior (“yes”) baseline, which ignores both the image *and* question (responding to every question with a “yes”).

Interestingly, the language-alone methods (per Q-type prior, BoW Q, LSTM Q)

that ignore the image perform surprisingly well, with BoW Q achieving 48.09% on open-ended (53.68% on multiple-choice) and LSTM Q achieving 48.76% on open-ended (54.75% on multiple-choice); both outperforming the nearest neighbor baseline (open-ended: 42.70%, multiple-choice: 48.49%). Our quantitative results and analyses suggest that this might be due to the language-model exploiting subtle statistical priors about the question types (e.g. “What color is the banana?” can be answered with “yellow” without looking at the image). For a detailed discussion of the subtle biases in the questions, please see [505].

The accuracy of our **best model** (deeper LSTM Q + norm I (Fig. 10), selected using VQA test-dev accuracies) on VQA test-standard is 58.16% (open-ended) / 63.09% (multiple-choice). We can see that our model is able to significantly outperform both the vision-alone and language-alone baselines. As a general trend, results on multiple-choice are better than open-ended. All methods are significantly worse than human performance.

Our VQA demo is available on CloudCV [20] – <http://vqa.cloudcv.org/>. This will be updated with newer models as we develop them.

To gain further insights into these results, we computed accuracies by question type in Table 3. Interestingly, for question types that require more reasoning, such as “Is the” or “How many”, the scene-level image features do not provide any additional information. However, for questions that can be answered using scene-level information, such as “What sport,” we do see an improvement. Similarly, for questions whose answer may be contained in a generic caption we see improvement, such as “What animal”. For all question types, the results are worse than human accuracies.

We also analyzed the accuracies of our best model (deeper LSTM Q + norm I) on a subset of questions with certain specific (ground truth) answers. In Fig. 11, we show the average accuracy of the model on questions with 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). We

can see that the model performs well for answers that are common visual objects such as “wii”, “tennis”, “bathroom” while the performance is somewhat underwhelming for counts (*e.g.*, “2”, “1”, “3”), and particularly poor for higher counts (*e.g.*, “5”, “6”, “10”, “8”, “7”).

In Fig. 12, we show the distribution of 50 most frequently predicted answers when the system is correct on the VQA validation set (plot is sorted by prediction frequency, not accuracy). In this analysis, “system is correct” implies that it has VQA accuracy 1.0 (see section 3.2 for accuracy metric). We can see that the frequent ground truth answers (*e.g.*, “yes”, “no”, “2”, “white”, “red”, “blue”, “1”, “green”) are more frequently predicted than others when the model is correct.

Table 3: Open-ended test-dev results for different question types on real images (Q+C is reported on val). Machine performance is reported using the bag-of-words representation for questions. Questions types are determined by the one or two words that start the question. The percentage of questions for each type is shown in parentheses. Last and second last columns respectively show the average human age and average degree of commonsense required to answer the questions (as reported by AMT workers), respectively. See text for details.

| Question Type | Open-Ended | | | | | | Human Age | Commonsense |
|---------------------|------------|-------|-------|-------|-------|-------------------------|-----------------------------|-------------|
| | K = 1000 | | | Human | | To Be Able To Answer | To Be Able To Answer (%) | |
| | Q | Q + I | Q + C | Q | Q + I | | | |
| what is (13.84) | 23.57 | 34.28 | 43.88 | 16.86 | 73.68 | 09.07 | | 27.52 |
| what color (08.98) | 33.37 | 43.53 | 48.61 | 28.71 | 86.06 | 06.60 | | 13.22 |
| what kind (02.49) | 27.78 | 42.72 | 43.88 | 19.10 | 70.11 | 10.55 | | 40.34 |
| what are (02.32) | 25.47 | 39.10 | 47.27 | 17.72 | 69.49 | 09.03 | | 28.72 |
| what type (01.78) | 27.68 | 42.62 | 44.32 | 19.53 | 70.65 | 11.04 | | 38.92 |
| is the (10.16) | 70.76 | 69.87 | 70.50 | 65.24 | 95.67 | 08.51 | | 30.30 |
| is this (08.26) | 70.34 | 70.79 | 71.54 | 63.35 | 95.43 | 10.13 | | 45.32 |
| how many (10.28) | 43.78 | 40.33 | 47.52 | 30.45 | 86.32 | 07.67 | | 15.93 |
| are (07.57) | 73.96 | 73.58 | 72.43 | 67.10 | 95.24 | 08.65 | | 30.63 |
| does (02.75) | 76.81 | 75.81 | 75.88 | 69.96 | 95.70 | 09.29 | | 38.97 |
| where (02.90) | 16.21 | 23.49 | 29.47 | 11.09 | 43.56 | 09.54 | | 36.51 |
| is there (03.60) | 86.50 | 86.37 | 85.88 | 72.48 | 96.43 | 08.25 | | 19.88 |
| why (01.20) | 16.24 | 13.94 | 14.54 | 11.80 | 21.50 | 11.18 | | 73.56 |
| which (01.21) | 29.50 | 34.83 | 40.84 | 25.64 | 67.44 | 09.27 | | 30.00 |
| do (01.15) | 77.73 | 79.31 | 74.63 | 71.33 | 95.44 | 09.23 | | 37.68 |
| what does (01.12) | 19.58 | 20.00 | 23.19 | 11.12 | 75.88 | 10.02 | | 33.27 |
| what time (00.67) | 8.35 | 14.00 | 18.28 | 07.64 | 58.98 | 09.81 | | 31.83 |
| who (00.77) | 19.75 | 20.43 | 27.28 | 14.69 | 56.93 | 09.49 | | 43.82 |
| what sport (00.81) | 37.96 | 81.12 | 93.87 | 17.86 | 95.59 | 08.07 | | 31.87 |
| what animal (00.53) | 23.12 | 59.70 | 71.02 | 17.67 | 92.51 | 06.75 | | 18.04 |
| what brand (00.36) | 40.13 | 36.84 | 32.19 | 25.34 | 80.95 | 12.50 | | 41.33 |

Finally, evaluating our best model (deeper LSTM Q + norm I) on the validation

questions for which we have age annotations (how old a human needs to be to answer the question correctly), we estimate that our model performs as well as a 4.74 year old child! The average age required on the same set of questions is 8.98. Evaluating the same model on the validation questions for which we have commonsense annotations (whether the question requires commonsense to answer it), we estimate that it has degree of commonsense of 17.35%. The average degree of commonsense required on same set of questions is 31.23%. Again, these estimates reflect the age and commonsense perceived by MTurk workers that would be required to answer the question. See the appendix for details.

We further analyzed the performance of the model for different age groups on the validation questions for which we have age annotations. In Fig. 13, we computed the average accuracy of the predictions made by the model for questions belonging to different age groups. Perhaps as expected, the accuracy of the model decreases as the age of the question increases (from 61.07% at 3 – 4 age group to 47.83% at 18+ age group).

In Fig. 14, we show the distribution of age of questions for different levels of accuracies achieved by our system on the validation questions for which we have age annotations. It is interesting to see that the relative proportions of different age groups is consistent across all accuracy bins with questions belonging to the age group 5-8 comprising the majority of the predictions which is expected because 5-8 is the most common age group in the dataset (see Fig. 9).

Table 4 shows the accuracy of different ablated versions of our best model (deeper LSTM Q + norm I) for both the open-ended and multiple-choice tasks on the VQA test-dev for real images. The different ablated versions are as follows –

1. **Without I Norm:** In this model, the activations from the last hidden layer of VGGNet [408] are not ℓ_2 -normalized. Comparing the accuracies in Table 4 and Table 2, we can see that ℓ_2 -normalization of image features boosts the

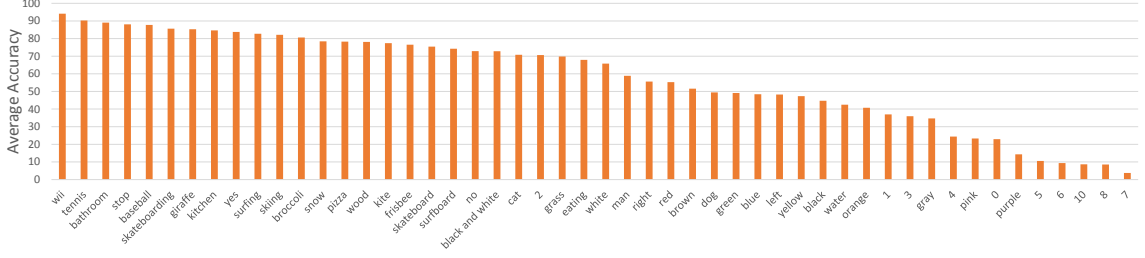


Figure 11: $\Pr(\text{system is correct} \mid \text{answer})$ for 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). System refers to our best model (deeper LSTM Q + norm I).

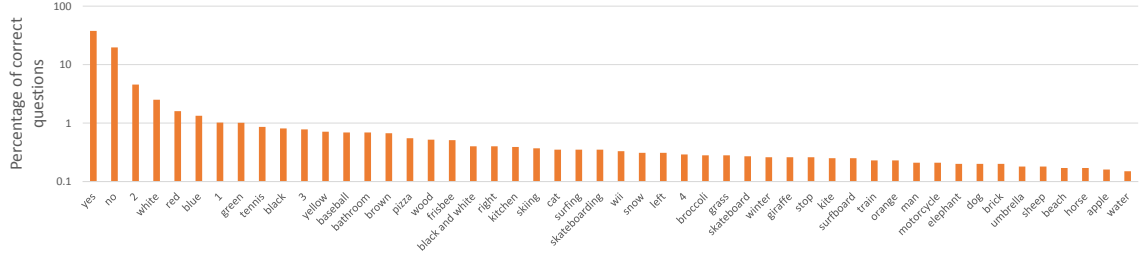


Figure 12: $\Pr(\text{answer} \mid \text{system is correct})$ for 50 most frequently predicted answers on the VQA validation set (plot is sorted by prediction frequency, not accuracy). System refers to our best model (deeper LSTM Q + norm I).

performance by 0.16% for open-ended task and by 0.24% for multiple-choice task.

2. **Concatenation:** In this model, the transformed image and LSTM embeddings are concatenated (instead of element-wise multiplied), resulting in doubling the number of parameters in the following fully-connected layer. Comparing the accuracies in Table 4 and Table 2, we can see that element-wise fusion performs better by 0.95% for open-ended task and by 1.24% for multiple-choice task.
3. **$K = 500$:** In this model, we use $K = 500$ most frequent answers as possible outputs. Comparing the accuracies in Table 4 and Table 2, we can see that $K = 1000$ performs better than $K = 500$ by 0.82% for open-ended task and by 1.92% for multiple-choice task.
4. **$K = 2000$:** In this model, we use $K = 2000$ most frequent answers as possible outputs. Comparing the accuracies in Table 4 and Table 2, we can see that K

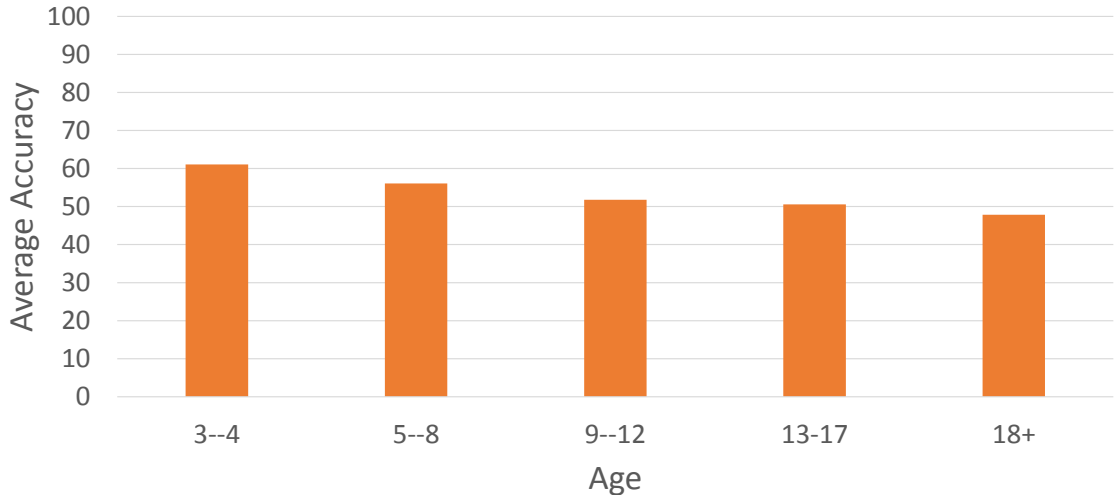


Figure 13: $\Pr(\text{system is correct} \mid \text{age of question})$ on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

= 2000 performs better than $K = 1000$ by 0.40% for open-ended task and by 1.16% for multiple-choice task.

5. **Truncated Q Vocab @ 5:** In this model, the input vocabulary to the embedding layer (which encodes the question words) consists of only those question words which occur at least 5 times in the training dataset, thus reducing the vocabulary size from 14770 (when all question words are used) to 5134 (65.24% reduction). Remaining question words are replaced with UNK (unknown) tokens. Comparing the accuracies in Table 4 and Table 2, we can see that truncating the question vocabulary @ 5 performs better than using all questions words by 0.24% for open-ended task and by 0.17% for multiple-choice task.
6. **Truncated Q Vocab @ 11:** In this model, the input vocabulary to the embedding layer (which encodes the question words) consists of only those question words which occur at least 11 times in the training dataset, thus reducing the vocabulary size from 14770 (when all question words are used) to 3561 (75.89%

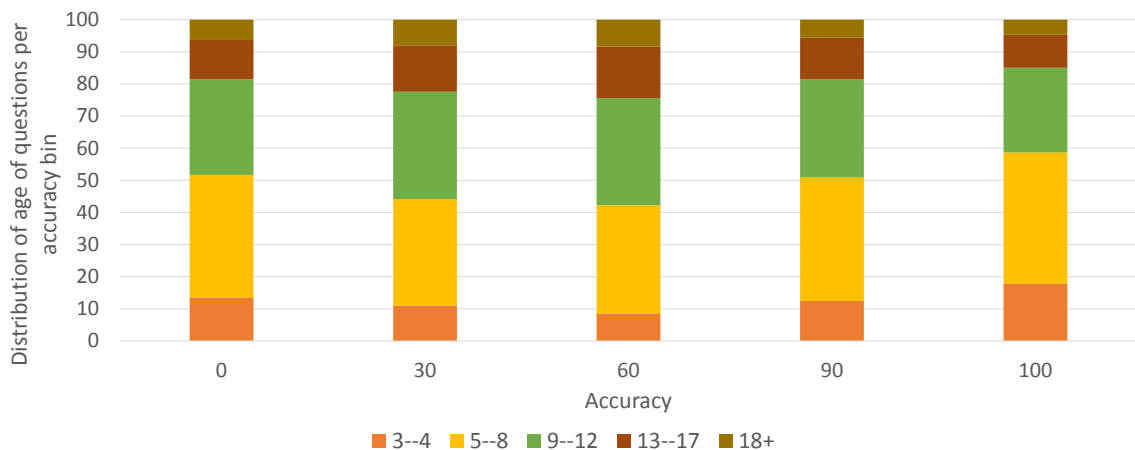


Figure 14: $\Pr(\text{age of question} \mid \text{system is correct})$ on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

reduction). Remaining question words are replaced with UNK (unknown) tokens. Comparing the accuracies in Table 4 and Table 2, we can see that truncating the question vocabulary @ 11 performs better than using all questions words by 0.06% for open-ended task and by 0.02% for multiple-choice task.

7. **Filtered Dataset:** We created a filtered version of the VQA train + val dataset in which we only keep the answers with subject confidence “yes”. Also, we keep only those questions for which at least 50% (5 out of 10) answers are annotated with subject confidence “yes”. The resulting filtered dataset consists of 344600 questions, compared to 369861 questions in the original dataset, leading to only 6.83% reduction in the size of the dataset. The filtered dataset has 8.77 answers per question on average. We did not filter the test set so that accuracies of the model trained on the filtered dataset can be compared with that of the model trained on the original dataset. The row “Filtered Dataset” in Table 4 shows the performance of the deeper LSTM Q + norm I model when trained on the filtered dataset. Comparing these accuracies with the corresponding accuracies in Table 2, we can see that the model trained on filtered version performs worse by 1.13% for open-ended task and by 1.88% for multiple-choice task.

Table 4: Accuracy of ablated versions of our best model (deeper LSTM Q + norm I) for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image. See text for details.

| | Open-Ended | | | | Multiple-Choice | | | |
|------------------------|------------|--------|--------|-------|-----------------|--------|--------|-------|
| | All | Yes/No | Number | Other | All | Yes/No | Number | Other |
| Without I Norm | 57.59 | 80.41 | 36.63 | 42.84 | 62.46 | 80.43 | 38.10 | 52.62 |
| Concatenation | 56.80 | 78.49 | 35.08 | 43.19 | 61.46 | 78.52 | 36.43 | 52.54 |
| K = 500 | 56.93 | 80.61 | 36.24 | 41.39 | 60.78 | 80.64 | 37.44 | 49.10 |
| K = 2000 | 58.15 | 80.56 | 37.04 | 43.79 | 63.86 | 80.59 | 38.97 | 55.20 |
| Truncated Q Vocab @ 5 | 57.99 | 80.67 | 36.99 | 43.38 | 62.87 | 80.71 | 38.22 | 53.20 |
| Truncated Q Vocab @ 11 | 57.81 | 80.42 | 36.97 | 43.22 | 62.72 | 80.45 | 38.30 | 53.09 |
| Filtered Dataset | 56.62 | 80.19 | 37.48 | 40.95 | 60.82 | 80.19 | 37.48 | 49.57 |

3.5 VQA Challenge and Workshop

We have set up an evaluation server³ where results may be uploaded for the test set and it returns an accuracy breakdown. We are organizing an annual challenge and workshop to facilitate systematic progress in this area; the first instance of the workshop was held at CVPR 2016⁴. We suggest that papers reporting results on the VQA dataset –

1. Report test-standard accuracies, which can be calculated using either of the non-test-dev phases, i.e., “test2015” or “Challenge test2015” on the following links: [[oe-real](#) | [oe-abstract](#) | [mc-real](#) | [mc-abstract](#)].
2. Compare their test-standard accuracies with those on the corresponding test2015 leaderboards [[oe-real-leaderboard](#) | [oe-abstract-leaderboard](#) | [mc-real-leaderboard](#) | [mc-abstract-leaderboard](#)].

For more details, please see the challenge page⁵. Screenshots of leaderboards for open-ended-real and multiple-choice-real are shown in Fig. 15. We also compare the test-standard accuracies of our best model (deeper LSTM Q + norm I) for both open-ended and multiple-choice tasks (real images) with other entries (as of October 28, 2016) on the corresponding leaderboards in Table 5.

³http://visualqa.org/challenge_2016.html

⁴http://www.visualqa.org/workshop_2016.html

⁵http://visualqa.org/challenge_2016.html

Table 5: Test-standard accuracy of our best model (deeper LSTM Q + norm I) compared to test-standard accuracies of other entries for the open-ended and multiple-choice tasks in the respective VQA Real Image Challenge leaderboards (as of October 28, 2016).

| | Open-Ended | | | | Multiple-Choice | | | |
|------------------------|------------|--------|--------|-------|-----------------|--------|--------|-------|
| | All | Yes/No | Number | Other | All | Yes/No | Number | Other |
| snubi-naverlabs | 60.60 | 82.23 | 38.22 | 46.99 | 64.95 | 82.25 | 39.56 | 55.68 |
| MM_PaloAlto | 60.36 | 80.43 | 36.82 | 48.33 | — | — | — | — |
| LV-NUS | 59.54 | 81.34 | 35.67 | 46.10 | 64.18 | 81.25 | 38.30 | 55.20 |
| ACVT_Adelaide | 59.44 | 81.07 | 37.12 | 45.83 | — | — | — | — |
| global_vision | 58.43 | 78.24 | 36.27 | 46.32 | — | — | — | — |
| deeper LSTM Q + norm I | 58.16 | 80.56 | 36.53 | 43.73 | 63.09 | 80.59 | 37.70 | 53.64 |
| iBOWIMG | — | — | — | — | 61.97 | 76.86 | 37.30 | 54.60 |

3.6 Conclusion and Discussion

In conclusion, we introduce the task of Visual Question Answering (VQA). Given an image and an open-ended, natural language question about the image, the task is to provide an accurate natural language answer. We provide a dataset containing over 250K images, 760K questions, and around 10M answers. We demonstrate the wide variety of questions and answers in our dataset, as well as the diverse set of AI capabilities in computer vision, natural language processing, and commonsense reasoning required to answer these questions accurately.

The questions we solicited from our human subjects were open-ended and not task-specific. For some application domains, it would be useful to collect task-specific questions. For instance, questions may be gathered from subjects who are visually impaired [51], or the questions could be focused on one specific domain (say sports). Bigham *et al.* [51] created an application that allows the visually impaired to capture images and ask open-ended questions that are answered by human subjects. Interestingly, these questions can rarely be answered using generic captions. Training on task-specific datasets may help enable practical VQA applications.

We believe VQA has the distinctive advantage of pushing the frontiers on “AI-complete” problems, while being amenable to automatic evaluation. Given the recent

Updated: 2016-04-17 (results migrated weekly from [CodaLab](#)).
For information about each test split please see the [challenge](#) page.

| | By Answer Type | | | Overall |
|---|----------------|--------|-------|---------|
| | Yes/No | Number | Other | |
| snubi-na-verlabs ^[3] | 82.23 | 38.22 | 46.99 | 60.6 |
| MM_PaloAlto ^[3] | 80.43 | 36.82 | 48.33 | 60.36 |
| LV-NUS ^[2] | 81.34 | 35.67 | 46.1 | 59.54 |
| ACVT_Adelaide ^[1] | 81.07 | 37.12 | 45.83 | 59.44 |
| global_vision ^[4] | 78.24 | 36.27 | 46.32 | 58.43 |
| vqteam-deeperLSTM_NormIzeCNN ^[7] | 80.96 | 36.53 | 43.73 | 58.16 |
| vqteam-istm_cnn ^[8] | 79.01 | 35.55 | 36.8 | 54.06 |
| vqteam-q_istm_alone ^[11] | 78.12 | 34.94 | 26.99 | 48.89 |
| vqteam-nearest_neighbor ^[9] | 71.73 | 24.31 | 22 | 42.73 |
| vqteam-prior_per_qtype ^[10] | 71.17 | 35.63 | 9.32 | 37.55 |
| vqteam-all_yes ^[6] | 70.53 | 0.43 | 1.26 | 29.72 |

Updated: 2016-04-17 (results migrated weekly from [CodaLab](#)).
For information about each test split please see the [challenge](#) page.

| | By Answer Type | | | Overall |
|---|----------------|--------|-------|---------|
| | Yes/No | Number | Other | |
| snubi-na-verlabs ^[3] | 82.25 | 39.56 | 55.68 | 64.95 |
| LV-NUS ^[1] | 81.25 | 38.3 | 55.2 | 64.18 |
| vqteam-deeperLSTM_NormIzeCNN ^[5] | 80.59 | 37.7 | 53.64 | 63.09 |
| iBOWIMG ^[2] | 76.86 | 37.3 | 54.6 | 61.97 |
| vqteam-istm_cnn ^[6] | 79.02 | 36.1 | 43.93 | 57.57 |
| vqteam-q_istm_alone ^[9] | 78.12 | 35.86 | 39.44 | 55.01 |
| vqteam-nearest_neighbor ^[7] | 71.75 | 25.81 | 34.09 | 48.75 |
| vqteam-prior_per_qtype ^[8] | 71.15 | 35.7 | 13.1 | 39.38 |
| vqteam-all_yes ^[4] | 70.53 | 0.43 | 1.26 | 29.72 |

Figure 15: Leaderboard showing test-standard accuracies for VQA Real Image Challenge (Open-Ended) on left and leaderboard showing test-standard accuracies for VQA Real Image Challenge (Multiple-Choice) on right (snapshot from October 28, 2016).

progress in the community, we believe the time is ripe to take on such an endeavor.

CHAPTER IV

ANALYZING THE BEHAVIOR OF VISUAL QUESTION ANSWERING MODELS

4.1 *Introduction*

After the release of our VQA dataset, a flurry of recent deep-learning based models have been proposed for VQA [27, 84, 495, 485, 208, 24, 465, 217, 280, 25, 400, 227, 150, 321, 197, 480, 483, 510, 381]. Curiously, the performance of most methods is clustered around 60-70% (compared to human performance of 83% on open-ended task and 91% on multiple-choice task) with a mere 5% gap between the top-9 entries on the VQA challenge 2016.¹ It seems clear that as a first step to understand these models, to meaningfully compare strengths and weaknesses of different models, to develop insights into their failure modes, and to identify the most fruitful directions for progress, it is crucial to develop techniques to understand the behavior of VQA models.

In this chapter, we develop novel techniques for characterizing the behavior of VQA models. As concrete instantiations, we analyze two VQA models ([279],[280]), one from each of the two major classes of VQA models – with-attention and without-attention. We also analyze the winning entry [150] of the VQA Challenge 2016.

4.2 *Behavior Analyses*

We analyze the behavior of VQA models along the following three dimensions –

Generalization to novel instances: We investigate whether the test instances that are incorrectly answered are the ones that are “novel” i.e., not similar to training

¹http://www.visualqa.org/challenge_2016.html

instances. The novelty of the test instances may be in two ways – 1) the test question-image (QI) pair is “novel”, i.e., too different from training QI pairs; and 2) the test QI pair is “familiar”, but the answer required at test time is “novel”, i.e., answers seen during training are different from what needs to be produced for the test QI pairs.

Complete question understanding: To investigate whether a VQA model is understanding the input question or not, we analyze whether the model ‘listens’ to only first few words of the question or the entire question, whether it ‘listens’ to only question (wh) words and nouns or all the words in the question.

Complete image understanding: The absence of a large gap between performance of language-alone and language + vision VQA models [27] provides evidence that current VQA models seem to be heavily reliant on the language model, perhaps not really understanding the image. In order to analyze this behavior, we investigate whether the predictions of the model change across images for a given question.

We present our behavioral analyses on the VQA dataset [27]. All the experimental results are reported on the VQA validation set using the following models trained on the VQA train set for the open-ended task –

CNN + LSTM based model without-attention (CNN+LSTM): We use the best performing model of [27] (code provided by [279]), which achieves an accuracy of 54.13% on the VQA validation set. It is a two channel model – one channel processes the image (using Convolutional Neural Network (CNN) to extract image features) and the other channel processes the question (using Long Short-Term Memory (LSTM) recurrent neural network to obtain question embedding). The image and question features obtained from the two channels are combined and passed through a fully connected (FC) layer to obtain a softmax distribution over the space of answers.

CNN + LSTM based model with-attention (ATT): We use the top-entry on the VQA challenge leaderboard (as of June 03, 2016) [280], which achieves an

accuracy of 57.02% on the VQA validation set.² This model jointly reasons about image and question attention, in a hierarchical fashion. The attended image and question features obtained from different levels of the hierarchy are combined and passed through a FC layer to obtain a softmax distribution over the space of answers.

VQA Challenge 2016 winning entry (MCB): This is the multimodal compact bilinear (mcb) pooling model from [150] which won the real image track of the VQA Challenge 2016. This model achieves an accuracy of 60.36% on the VQA validation set.³ In this model, multimodal compact bilinear pooling is used to predict attention over image features and also to combine the attended image features with the question features. These combined features are passed through a FC layer to obtain a softmax distribution over the space of answers.

4.2.1 Generalization to novel instances

Do VQA models make mistakes because test instances are too different from training ones? To analyze the first type of novelty (the test QI pair is novel), we measure the correlation between test accuracy and distance of test QI pairs from its k nearest neighbor (k-NN) training QI pairs. For each test QI pair we find its k-NNs in the training set and compute the average distance between the test QI pair and its k-NNs. The k-NNs are computed in the space of combined image + question embedding (just before passing through FC layer) for all the three models (using euclidean distance metric for the CNN+LSTM model and cosine distance metric for the ATT and MCB models).

The correlation between accuracy and average distance is significant (-0.41 at $k=50$ ⁴ for the CNN+LSTM model and -0.42 at $k=15$ ⁵ for the ATT model). A high negative correlation value tells that the model is less likely to predict correct answers

²Code available at <https://github.com/jiasenlu/HieCoAttenVQA>

³Code available at <https://github.com/akirafukui/vqa-mcb>

⁴ $k=50$ leads to highest correlation

⁵ $k=15$ leads to highest correlation

for test QI pairs which are not very similar to training QI pairs, suggesting that the model is not very good at generalizing to novel test QI pairs. The correlation between accuracy and average distance is not significant for the MCB model (-0.14 at $k=1$ ⁶) suggesting that MCB is better at generalizing to novel test QI pairs.

We also found that 67.5% of mistakes made by the CNN+LSTM model *can be successfully predicted* by checking distance of test QI pair from its k-NN training QI pairs (66.7% for the ATT model, 55.08% for the MCB model). Thus, this analysis not only exposes a reason for mistakes made by VQA models, but also allows us to build human-like models that can predict their own oncoming failures, and potentially refuse to answer questions that are ‘too different’ from ones seen in past.

To analyze the second type of novelty (the answer required at test time is not familiar), we compute the correlation between test accuracy and the average distance of the test ground truth (GT) answer with GT answers of its k-NN training QI pairs. The distance between answers is computed in the space of average Word2Vec [301] vectors of answers. This correlation turns out to be quite high (-0.62) for both CNN+LSTM and ATT models and significant (-0.47) for the MCB model. A high negative correlation value tells that the model tends to regurgitate answers seen during training.

These distance features are also good at predicting failures – 74.19% of failures can be predicted by checking distance of test GT answer with GT answers of its k-NN training QI pairs for CNN+LSTM model (75.41% for the ATT model, 70.17% for the MCB model). Note that unlike the previous analysis, this analysis only explains failures but cannot be used to predict failures (since it uses GT labels). See Fig. 16 for qualitative examples.

From Fig. 16 (row1) we can see that the test QI pair is semantically quite different from its k-NN training QI pairs ({1st, 2nd, 3rd}-NN distances are {15.05, 15.13,

⁶ $k=1$ leads to highest correlation



Figure 16: Examples from test set where the CNN+LSTM model makes mistakes and their corresponding nearest neighbor training instances. See appendix for more examples.

15.17}, which are higher than the corresponding distances averaged across all success cases: {8.74, 9.23, 9.50.}), explaining the mistake. Row2 shows an example where the model has seen the same question in the training set (test QI pair is semantically similar to training QI pairs) but, since it has not seen “green cone” for training instances (answers seen during training are different from what needs to be produced for the test QI pair), it is unable to answer the test QI pair correctly. This shows that current models lack compositionality: the ability to combine the concepts of “cone” and “green” (both of which have been seen in training set) to answer “green cone” for the test QI pair. This compositionality is desirable and central to intelligence.

4.2.2 Complete question understanding

We feed partial questions of increasing lengths (from 0-100% of question from left to right). We then compute what percentage of responses do not change when more and more words are fed.

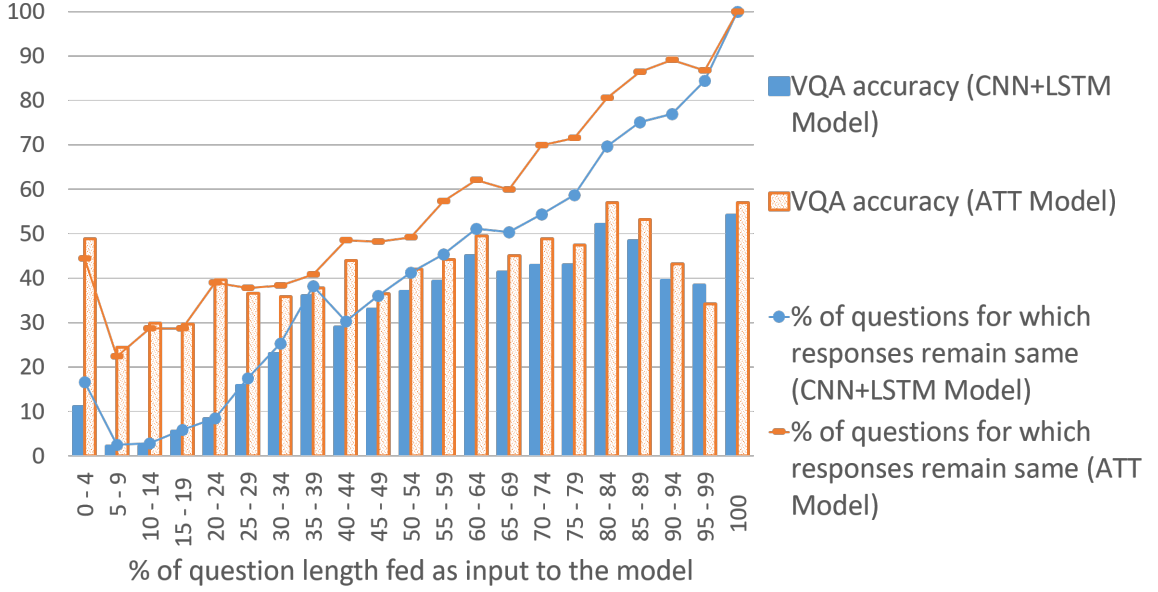


Figure 17: X-axis shows length of partial question (in %) fed as input. Y-axis shows percentage of questions for which responses of these partial questions are the same as full questions and VQA accuracy of partial questions.

Fig. 17 shows the test accuracy and percentage of questions for which responses remain same (compared to entire question) as a function of partial question length. We can see that for 40% of the questions, the CNN+LSTM model seems to have converged on a predicted answer after ‘listening’ to just half the question. This shows that the model is listening to first few words of the question more than the words towards the end. Also, the model has 68% of the final accuracy (54%) when making predictions based on half the original question. When making predictions just based on the image, the accuracy of the model is 24%. The ATT model seems to have converged on a predicted answer after listening to just half the question more often (49% of the time), achieving 74% of the final accuracy (57%). The MCB model converges on a predicted answer after listening to just half the question 45% of the time, achieving 67% of the final accuracy (60%). See Fig. 18 for qualitative examples.

We also analyze the change in responses of the model’s predictions (see Fig. 19),

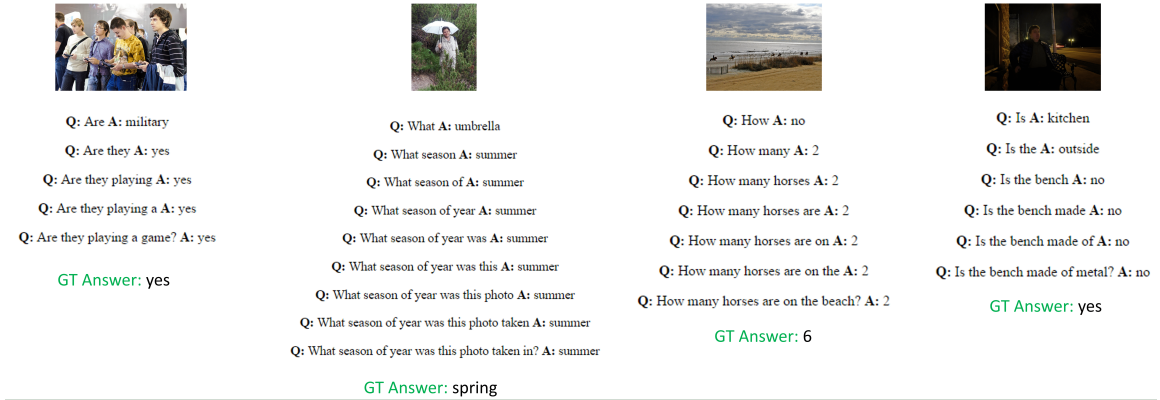


Figure 18: Examples where the CNN+LSTM model does not change its answer after first few question words. On doing so, it is correct for some cases (the extreme left example) and incorrect for other cases (the remaining three examples). See appendix for more examples.

when words of a particular part-of-the-speech (POS) tag are dropped from the question. The experimental results indicate that wh-words effect the model’s decisions the most (most of the responses get changed on dropping these words from the question), and that pronouns effect the model’s decisions the least.

4.2.3 Complete image understanding

Does a VQA model really ‘look’ at the image? To analyze this, we compute the percentage of the time (say X) the response does not change across images (e.g.,, answer for all images is “2”) for a given question (e.g., “How many zebras?”) and plot histogram of X across questions (see Fig. 20). We do this analysis for questions occurring for atleast 25 images in the VQA validation set, resulting in total 263 questions. The cumulative plot indicates that for 56% questions, the CNN+LSTM model outputs the same answer for at least half the images. This is fairly high, suggesting that the model is picking the same answer no matter what the image is. Promisingly, the ATT and MCB models (that do not work with a holistic entire-image representation and purportedly pay attention to specific spatial regions in an

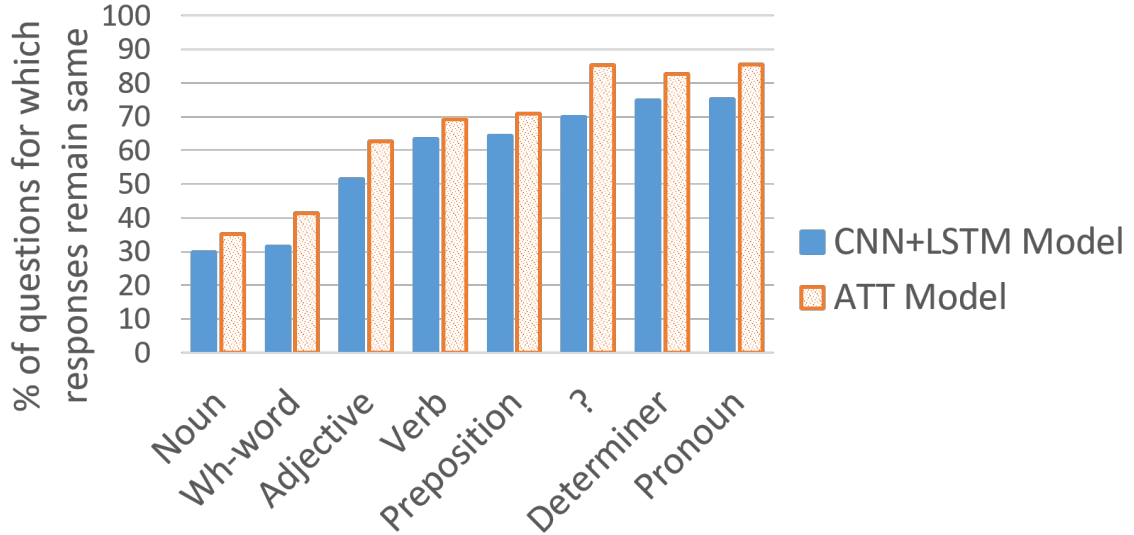


Figure 19: Percentage of questions for which responses remain same (compared to entire question) as a function of POS tags dropped from the question.

image) produce the same response for at least half the images for fewer questions (42% for the ATT model, 40% for the MCB model).

Interestingly, the average accuracy (see the VQA accuracy plots in Fig. 20) for questions for which the models produce same response for $>50\%$ and $<55\%$ of the images is 56% for the CNN+LSTM model (60% for the ATT model, 73% for the MCB model) which is more than the respective average accuracy on the entire VQA validation set (54.13% for the CNN+LSTM model, 57.02% for the ATT model, 60.36% for the MCB model). Thus, producing the same response across images seems to be statistically favorable. Fig. 21 shows examples where the CNN+LSTM model predicts the same response across images for a given question. The first row shows examples where the model makes errors on several images by predicting the same answer for all images. The second row shows examples where the model is always correct even if it predicts the same answer across images. This is so because questions such as “*What covers the ground?*” are asked for an image in the VQA dataset only when ground is covered with snow (because subjects were looking at the image while asking questions about it). Thus, this analysis exposes label biases in the

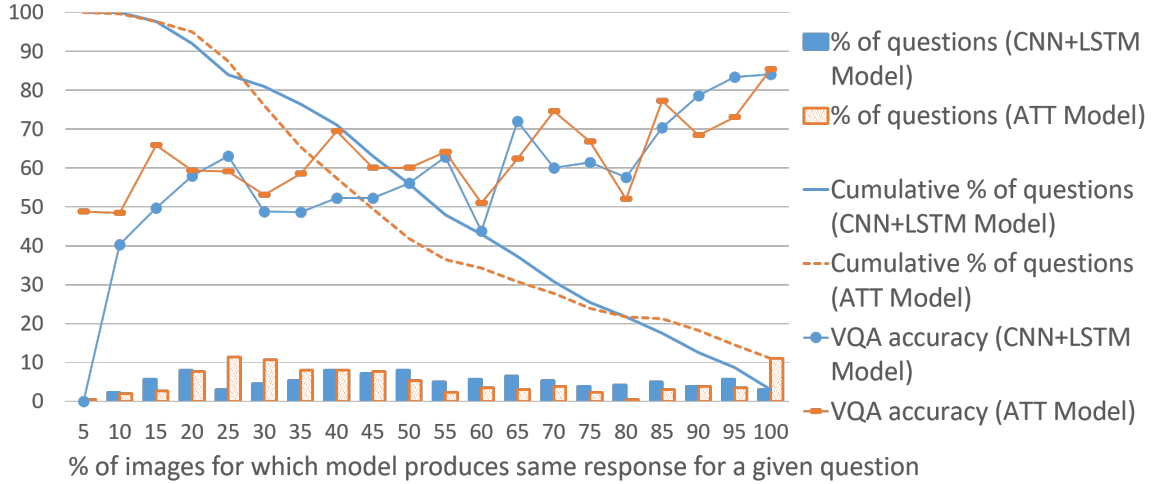


Figure 20: Histogram of percentage of images for which model produces same answer for a given question and its comparison with test accuracy. The cumulative plot shows the % of questions for which model produces same answer for *atleast* x % of images.

dataset. Label biases (in particular, for “yes/no” questions) have also been reported in [505].

Correct Response

Predicted A: 2



Incorrect Responses

Q: How many zebras

Predicted A: 2



Predicted A: 2



Predicted A: 2



All Correct Responses

Q: What covers the ground

Predicted A: snow



Predicted A: snow



Predicted A: snow



Predicted A: snow



Predicted A: snow



Figure 21: Examples where the predicted answers do not change across images for a given question. See appendix for more examples.

4.3 Conclusion

We develop novel techniques to characterize the behavior of VQA models, as a first step towards understanding these models, meaningfully comparing the strengths and

weaknesses of different models, developing insights into their failure modes, and identifying the most fruitful directions for progress. Our behavior analysis reveals that despite recent progress, today’s VQA models are “myopic” (tend to fail on sufficiently novel instances), often “jump to conclusions” (converge on a predicted answer after ‘listening’ to just half the question), and are “stubborn” (do not change their answers across images), with attention based models being less “stubborn” than non-attention based models.

As a final thought, we note that the somewhat pathological behaviors exposed in the paper are in some sense “correct” given the model architectures and the dataset being trained on. Ignoring optimization error, the maximum-likelihood training objective is clearly intended to capture statistics of the dataset. Our motive is simply to better understand current generation models via their behaviors, and use these observations to guide future choices – do we need novel model classes? or dataset with different biases? etc. Finally, it should be clear that our use of anthropomorphic adjectives such as “stubborn”, “myopic” etc. is purely for pedagogical reasons – to easily communicate our observations to our readers. No claims are being made about today’s VQA models being human-like.

CHAPTER V

OVERCOMING PRIORS IN VISUAL QUESTION ANSWERING

5.1 Visual Question Answering under Changing Priors (VQA-CP)

5.1.1 Introduction

In Chapter 4 we saw that today’s VQA models are heavily driven by superficial correlations in the training data and lack sufficient visual grounding. Similar findings have been reported in other works as well [9, 505, 168, 211]. It seems that when faced with a difficult learning problem, models typically resort to latching onto the language priors in the training data to the point of ignoring the image – *e.g.*, overwhelmingly replying to ‘*how many X?*’ questions with ‘2’ (irrespective of X), ‘*what color is ... ?*’ with ‘*white*’, ‘*is the ... ?*’ with ‘*yes*’.

One reason for this emergent dissatisfactory behavior is the fundamentally problematic nature of IID train-test splits *in the presence of strong priors*. As a result, models that intrinsically memorize biases in the training data demonstrate acceptable performance on the test set. This is problematic for benchmarking progress in VQA because it becomes unclear what the source of the improvements is – if models have learned to ground concepts in images or they are driven by memorizing priors in training data.

To help disentangle these factors, we present new splits of the VQA v1 [27] and VQA v2 [168] datasets, called **Visual Question Answering under Changing Priors** (**VQA-CP v1** and **VQA-CP v2** respectively). These new splits are created by re-organizing the train and val splits of the respective VQA datasets in such a way

that the distribution of answers per question type (*‘how many’, ‘what color is’, etc.*) is by design *different* in the test split compared to the train split (Section 5.1.2). One important thing to note: we do not change the distribution of the underlying perceptual signals – the images – between train and test. Generalization across different domains of images (*e.g.* COCO images *vs.* web cam images) is an active research area and not the focus of this work. We change the distribution of *answers for each question type* between train and test. Our hypothesis is that it is reasonable to expect models that are answering questions for the ‘right reasons’ (image grounding) to recognize, for instance, ‘*black*’ color at test time even though ‘*white*’ is the most popular answer for ‘*What color is the ... ?*’ questions in the train set.

To demonstrate the difficulty of our VQA-CP splits, we report the performance of several existing VQA models [279, 24, 495, 150] on these splits. Our key finding is that the performance of *all tested existing* models drops significantly when trained and evaluated on the new splits compared to the original splits (Section 5.1.3). This finding provides further confirmation and a novel insight to the growing evidence in literature on the behavior of VQA models [9, 505, 168, 211].

5.1.2 VQA-CP : Dataset Creation and Analysis

The VQA-CP v1 and VQA-CP v2 splits are created such that the distribution of answers per question type (*‘how many’, ‘what color is’, etc.*) is different in the test data compared to the training data. These splits are created by re-organizing the training and validation splits of the VQA v1 [27] and VQA v2 [168] datasets respectively ¹, using the following procedure:

Question Grouping: Questions having the same question type (first few words of the question – ‘*What color is the*’, ‘*What room is*’, etc.) and the same ground truth answer are grouped together. For instance, {‘*What color is the dog?*’, ‘*white*’}

¹We can not use the test splits from VQA datasets because creation of VQA-CP splits requires access to answer annotations, which are not publicly available on the test sets.

and $\{ \text{'What color is the plate?'}, \text{'white'} \}$ are grouped together whereas $\{ \text{'What color is the dog?'}, \text{'black'} \}$ is put in a different group. This grouping is done after merging the QA pairs from the VQA train and val splits. We use the question types provided in the VQA datasets.

Greedy Re-splitting: A greedy approach is used to redistribute data points (image, question, answer) to the VQA-CP train and test splits so as to maximize the coverage of the VQA-CP test concepts in the VQA-CP train split while making sure that questions with the same question type and the same ground truth answer are not repeated between test and train splits. In this procedure, we loop through all the groups created above, and in every iteration, we add the current group to the VQA-CP test split unless the group has already been assigned to the VQA-CP train split. We always maintain a set of concepts² belonging to the groups in the VQA-CP test split that have not yet been covered by the groups in the VQA-CP train split. We then pick the group that covers majority of the concepts in the set, from the groups that have not yet been assigned to either split and add that group to the VQA-CP train split. We stop when the test split has about 1/3rd the dataset and add the remaining groups (not yet assigned to either split) to the train split.

The above approach results in 98.04% coverage of test question concepts (set of all unique words in questions after removing stop words – *'is', 'are', 'the', etc.*) in the train split for VQA-CP v1 (99.01% for VQA-CP v2), and 95.07% coverage of test answers by the train split’s top 1000 answers for VQA-CP v1 (95.72% for VQA-CP v2). VQA-CP v1 train consists of $\sim 118\text{K}$ images, $\sim 245\text{K}$ questions and $\sim 2.5\text{M}$ answers ($\sim 121\text{K}$ images, $\sim 438\text{K}$ questions and $\sim 4.4\text{M}$ answers for VQA-CP v2 train). VQA-CP v1 test consists of $\sim 87\text{K}$ images, $\sim 125\text{K}$ questions and $\sim 1.3\text{M}$ answers ($\sim 98\text{K}$ images, $\sim 220\text{K}$ questions and $\sim 2.2\text{M}$ answers for VQA-CP v2 test).

²For a given group, concepts are the set of all unique words present in the question type and the ground truth answer belonging to that group.

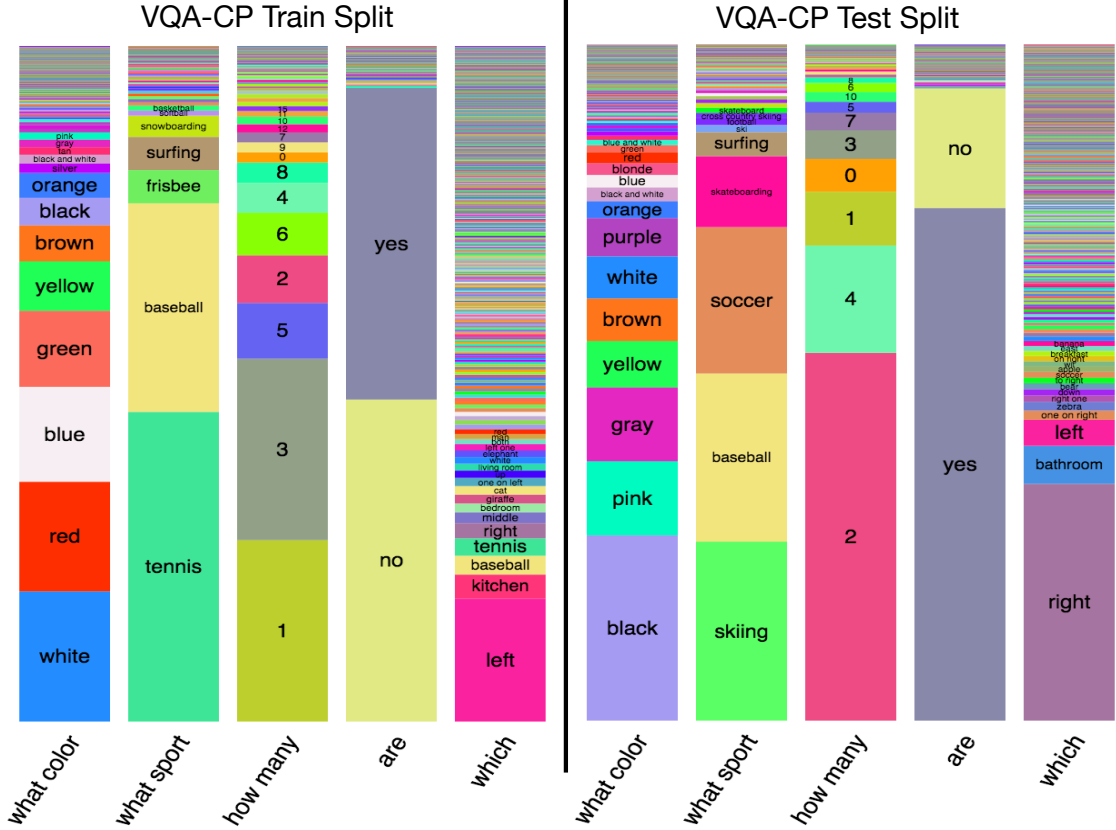


Figure 22: Distribution of answers per question type vary significantly between VQA-CP v1 train (left) and test (right) splits. For instance, ‘white’ and ‘red’ are commonly seen answers in train for ‘What color’, where as ‘black’ is the most frequent answer in test. These have been computed for a random sample of 60K questions.

Fig. 22 shows the distribution of answers for several question types such as ‘what color’, ‘what sport’, ‘how many’, etc. for the train (left) and test (right) splits of the VQA-CP v1 dataset (see appendix for this analysis of the VQA-CP v2 dataset). We can see that the distributions of answers for a given question type is significantly different. For instance, ‘tennis’ is the most frequent answer for the question type ‘what sport’ in VQA-CP v1 train split whereas ‘skiing’ is the most frequent answer for the same question type in VQA-CP v1 test split. However, for VQA v1 dataset, the distribution for a given question type is similar across train and val splits [27] (for instance, ‘tennis’ is the most frequent answer for both the train and val splits). In the VQA-CP v1 splits, similar differences can be seen for other question types as

Table 6: We compare the performance of existing VQA models on VQA-CP v1 test splits (when trained on VQA-CP v1 train splits) to their performance on VQA v1 val splits (when trained on VQA v1 train splits). We find that the performance of all tested existing models degrades significantly in the new Changing Priors setting compared to the original VQA setting.

| Model | Dataset | Overall | Yes/No | Number | Other |
|-------------------------|-----------|---------|--------|--------|-------|
| per Q-type prior [27] | VQA v1 | 35.13 | 71.31 | 31.93 | 08.86 |
| | VQA-CP v1 | 08.39 | 14.70 | 08.34 | 02.14 |
| d-LSTM Q [27] | VQA v1 | 48.23 | 79.05 | 33.70 | 28.81 |
| | VQA-CP v1 | 20.16 | 35.72 | 11.07 | 08.34 |
| d-LSTM Q + norm I [279] | VQA v1 | 54.40 | 79.82 | 33.87 | 40.54 |
| | VQA-CP v1 | 23.51 | 34.53 | 11.40 | 17.42 |
| NMN [24] | VQA v1 | 54.83 | 80.39 | 33.45 | 41.07 |
| | VQA-CP v1 | 29.64 | 38.85 | 11.23 | 27.88 |
| SAN [495] | VQA v1 | 55.86 | 78.54 | 33.46 | 44.51 |
| | VQA-CP v1 | 26.88 | 35.34 | 11.34 | 24.70 |
| MCB [150] | VQA v1 | 60.97 | 81.62 | 34.56 | 52.16 |
| | VQA-CP v1 | 34.39 | 37.96 | 11.80 | 39.90 |

well – ‘are’, ‘which’.

5.1.3 Benchmarking VQA Models on VQA-CP

To demonstrate the difficulty of our VQA-CP splits, we report the performance of the following baselines and existing VQA models when trained on VQA-CP v1 and VQA-CP v2 train splits and evaluated on the corresponding test splits. We compare this with their performance when trained on VQA v1 and VQA v2 train splits and evaluated on the corresponding val splits. Results are presented in Tables 5.1.2 and 5.1.2.

per Q-type prior [27]: Predicting the most popular training answer for the corresponding question type (e.g., ‘tennis’ for ‘What sport ...?’ questions) ³.

Deeper LSTM Question (d-LSTM Q) [27]: Predicting the answer using question

³Note that, ideally the performance of this baseline on VQA-CP test set should be zero because the answers, given the question type, are different in test and train. But, due to some inter-human disagreement in the datasets, the performance is slightly higher (Tables 5.1.2 and 5.1.2).

Table 7: We compare the performance of existing VQA models on VQA-CP v2 test splits (when trained on VQA-CP v2 train splits) to their performance on VQA v2 val splits (when trained on VQA v2 train splits). We find that the performance of all tested existing models degrades significantly in the new Changing Priors setting compared to the original VQA setting.

| Model | Dataset | Overall | Yes/No | Number | Other |
|-------------------------|-----------|---------|--------|--------|-------|
| per Q-type prior [27] | VQA v2 | 32.06 | 64.42 | 26.95 | 08.76 |
| | VQA-CP v2 | 08.76 | 19.36 | 11.70 | 02.39 |
| d-LSTM Q [27] | VQA v2 | 43.01 | 67.95 | 30.97 | 27.20 |
| | VQA-CP v2 | 15.95 | 35.09 | 11.63 | 07.11 |
| d-LSTM Q + norm I [279] | VQA v2 | 51.61 | 73.06 | 34.41 | 39.85 |
| | VQA-CP v2 | 19.73 | 34.25 | 11.39 | 14.41 |
| NMN [24] | VQA v2 | 51.62 | 73.38 | 33.23 | 39.93 |
| | VQA-CP v2 | 27.47 | 38.94 | 11.92 | 25.72 |
| SAN [495] | VQA v2 | 52.02 | 68.89 | 34.55 | 43.80 |
| | VQA-CP v2 | 24.96 | 38.35 | 11.14 | 21.74 |
| MCB [150] | VQA v2 | 59.71 | 77.91 | 37.47 | 51.76 |
| | VQA-CP v2 | 36.33 | 41.01 | 11.96 | 40.57 |

alone (“blind” model).

Deeper LSTM Question + normalized Image (d-LSTM Q + norm I) [27]:

The baseline VQA model.

Neural Module Networks (NMN) [24]: The model designed to be compositional in nature.

Stacked Attention Networks (SAN) [495]: One of the widely used models for VQA.

Multimodal Compact Bilinear Pooling (MCB) [150]: The winner of the VQA Challenge (on real image) 2016.

Brief descriptions of all of these models are in appendix.

From Tables 5.1.2 and 5.1.2, we can see that the performance of all tested existing VQA models drops significantly in the VQA-CP setting compared to the original VQA setting. Note that even though the NMN architecture is compositional by design, their performance degrades on the VQA-CP datasets. We posit this may

be because they use an additional LSTM encoding of the question to encode priors in the dataset. Also note that the d-LSTM Q + norm I model suffers the largest drop in overall performance compared to other VQA models, perhaps because other models have more powerful visual processing (for instance, attention on images). Another interesting observation from Tables 5.1.2 and 5.1.2 is that the ranking of the models based on overall performance changes from VQA to VQA-CP. For VQA, SAN outperforms NMN, whereas for VQA-CP, NMN outperforms SAN. For a brief discussion on trends for different question types, please see appendix.

5.1.4 Conclusion

In conclusion, we propose a new setting for VQA (VQA under Changing Priors (VQA-CP)) where, for every question type, train and test sets have different prior distributions of answers. We introduce novel splits of the existing VQA v1 and VQA v2 datasets to stress test models under changing priors. Quatitative evaluation of several existing VQA models on these new splits shows that the performance of all tested existing models drops significantly in the proposed Changing Priors setting compared to the existing setting where the train and test distributions of answers given the question type are similar. This finding provides further confirmation that today’s VQA models are largely driven by language priors in the training data and lack sufficient image grounding. Thus, the proposed splits can serve as benchmarks to evaluate the degree of visual groundedness in VQA models.

5.2 *Grounded Visual Question Answering (GVQA)*

5.2.1 Introduction

In this section, we propose a novel **Grounded Visual Question Answering (GVQA)** model that contains inductive biases and restrictions in the architecture specifically designed to prevent it from ‘cheating’ by primarily relying on priors in the training data (Section 5.2.2). GVQA is motivated by the intuition that questions in VQA

provide two key pieces of information:

- (1) What should be recognized? Or what visual concepts in the image need to be reasoned about to answer the question (*e.g.*, ‘*What color is the plate?*’ requires looking at the plate in the image),
- (2) What should be said? Or what is the space of plausible answers (*e.g.*, ‘*What color ... ?*’ questions need to be answered with names of colors).

Our hypothesis is that models that do not explicitly differentiate between these two roles – which is the case for most existing models in literature – tend to confuse these two signals. They end up learning from question-answer pairs that a plausible color of a plate is white, and at test time, rely on this correlation more so than the specific plate in the image the question is about. GVQA explicitly disentangles the visual concept recognition from the answer space prediction.

GVQA is built off of an existing VQA model – Stacked Attention Networks (SAN) [495]. Our experiments demonstrate that GVQA significantly outperforms SAN on both VQA-CP v1 and VQA-CP v2 datasets (Section 5.2.3). Interestingly, it also outperforms more powerful VQA models such as Multimodal Compact Bilinear Pooling (MCB) [150] in several cases (Section 5.2.3). We also show that GVQA offers strengths complementary to SAN when trained and evaluated on the original VQA v1 and VQA v2 datasets (Section 5.2.5). Finally, GVQA is more transparent than existing VQA models, in that it produces interpretable intermediate outputs unlike most existing VQA models (Section 5.2.6).

5.2.2 GVQA model

We now introduce our Grounded Visual Question Answering model (GVQA). While previous VQA approaches directly map Image-Question tuples (I, Q) to Answers (A) , GVQA breaks down the task of VQA into two steps: **Look** - locate the object / image patch needed to answer the question and recognize the visual concepts in the patch,

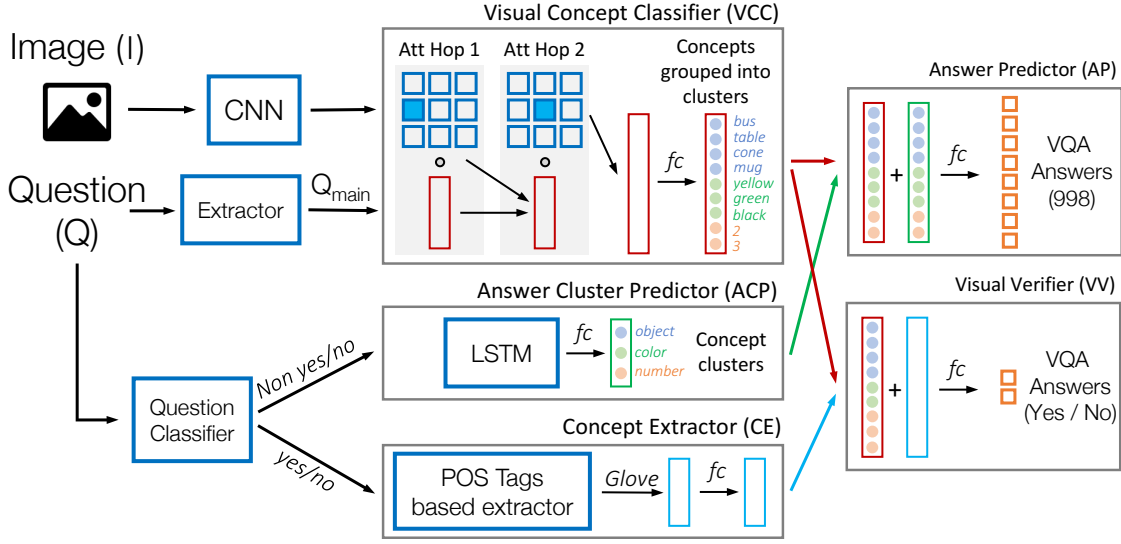


Figure 23: The proposed Grounded Visual Question Answering (GVQA) model.

and **Answer** - identify the space of plausible answers from the question and return the appropriate visual concept from the set of recognized visual concepts by taking into account which concepts are plausible. For instance, when GVQA is asked ‘*What color is the dog?*’, it identifies that the answer should be a color name, locates the patch in the image corresponding to dog, recognizes various visual concepts such as ‘*black*’, ‘*dog*’, ‘*furry*’, and finally outputs the concept ‘*black*’ because it is the recognized concept corresponding to color. Another novelty in GVQA is that it treats answering yes/no questions as a visual verification task, i.e., it verifies the visual presence/absence of the concept mentioned in the question. For instance, when GVQA is asked ‘*Is the person wearing shorts?*’, it identifies that the concept whose visual presence needs to be verified is ‘*shorts*’ and answers ‘*yes*’ or ‘*no*’ depending on whether it recognizes shorts or not in the image (specifically, on the patch corresponding to ‘*person*’).

GVQA is depicted in Figure 23. Given a question and an image, the question first goes through the *Question Classifier* and gets classified into yes/no or non yes/no. For non yes/no questions, the GVQA components that get activated are – 1) *Visual*

Concept Classifier (VCC) which takes as input the image features extracted from *CNN* and Q_{main} given by the question *Extractor*, 2) *Answer Cluster Predictor (ACP)* whose input is the entire question. The outputs of *VCC* and *ACP* are fed to the *Answer Predictor (AP)* which produces the answer. For yes/no questions, the GVQA components that get activated are – 1) *VCC* (similarly to non yes/no), 2) *Concept Extractor (CE)* whose input is the entire question. The outputs of *VCC* and *CE* are fed to the *Visual Verifier (VV)* which predicts ‘yes’ or ‘no’. We present the details of each component below.

Visual Concept Classifier (VCC) is responsible for locating the image patch that is needed to answer the question, as well as producing a set of visual concepts relevant to the located patch. E.g., given ‘*What is the color of the bus next to the car?*’, the VCC is responsible for attending on the bus region and then outputting a set of concepts such as ‘*bus*’ and attributes such as its color, count, etc. It consists of a 2-hop attention module based off of Stacked Attention Networks ([495]) followed by a stack of binary concept classifiers. The image is fed to the attention module in the form of activations of the last pooling layer of VGG-Net [410]. To prevent the memorization of answer priors per question type, the question is first passed through a language *Extractor*, a simple rule that outputs the string (called Q_{main}) after removing the question type substring (eg. ‘*What kind of*’). Q_{main} is embedded using an LSTM and then fed into the attention module. The multi hop attention produces a weighted linear combination of the image region features from VGG-Net, with weights corresponding to the degree of attention for that region. This is followed by a set of fully connected (FC) layers and a stack of ~ 2000 binary concept classifiers that cover $\sim 95\%$ of the concepts seen in train. VCC is trained with a binary logistic loss for every concept.

The set of VCC concepts is constructed by extracting objects and attributes, pertinent to the answer, from training QA pairs and retaining the most frequent ones.

Object concepts are then grouped into a single group where as attribute concepts are clustered into multiple small groups using K-means clustering in Glove embedding space [346], for a total of C clusters.⁴ Concept clustering is required for the purpose of generating negative samples required to train the concept classifiers (for a concept classifier, positive samples are those which contain that concept either in the question or the answer). Since the question does not indicate objects and attributes absent in the image, negative data is generated using the following assumptions: (1) the attended image patch required to answer a question has at most one dominant object in it (2) every object has at most one dominant attribute from each attribute category (e.g., if the color of a bus is red, it can be used as a negative example for all other colors). Given these assumptions, when a concept in a cluster is treated as positive, all other concepts in that cluster are treated as negatives. Note that only a subset of all concept clusters are activated for each question during training, and only these activated clusters contribute to the loss.

Question Classifier classifies the input question Q into 2 categories: Yes-No and non Yes-No using a Glove embedding layer, an LSTM and FC layers. Yes-No questions feed into the CE and the rest feed into the ACP.

Answer Cluster Predictor (ACP) identifies the *type* of the expected answer (e.g. object name, color, number, *etc.*). It is only activated for non yes/no questions. It consists of a Glove embedding layer and an LSTM, followed by FC layers that classify questions into one of the C clusters. The clusters for ACP are created by K-means clustering on (1000) answer classes by embedding each answer in Glove space.⁵

Concept Extractor (CE) extracts question concepts from yes/no questions

⁴We use $C = 50$ because it gives better clusters than other values. Also, agglomerative clustering results in similar performance as K-means. More details in appendix.

⁵We first create the clusters for ACP using the answer classes. We then create the clusters for VCC by assigning each VCC concept to one of these ACP clusters using Euclidean distance in Glove embedding space.

whose visual presence needs to be verified in the image, using a POS tag based extraction system⁶. E.g., for *‘Is the cone green?’*, we extract *‘green’*. The extracted concept is embedded in Glove space followed by FC layers to transform this embedding to the same space as the VCC concepts so that they can be combined by VV. Please see the description of VV below.

Answer Predictor (AP): Given a set of visual concepts predicted by the VCC, and a concept category predicted by the ACP, the AP’s role is to predict the answer. ACP categories correspond to VCC concept clusters (see ACP’s and VCC’s output classes in Fig. 23. The colors denote the correspondence). Given this alignment, the output of the ACP can be easily mapped into a vector with the same dimensions as the VCC output by simply copying ACP dimensions into positions pertaining to the respective VCC cluster dimensions. The resulting ACP embedding is added element-wise to the VCC embedding followed by FC layers and a softmax activation, yielding a distribution over 998 VQA answer categories (top 1000 training answers minus *‘yes’* and *‘no’*).

Visual Verifier (VV): Given a set of visual concepts predicted by the VCC and the embedding of the concept whose visual presence needs to be verified (given by CE), the VV’s role is to verify the presence/absence of the concept in VCC’s predictions. Specifically, the CE embedding is added element-wise to the VCC embedding followed by FC layers and a softmax activation, yielding a distribution over two categories – *‘yes’* and *‘no’*.

Model Training and Testing: We first train VCC and ACP on the train split using the cluster labels (for ACP) and visual concept labels (for VCC)⁷. The inputs to Answer Predictor (and Visual Verifier) are the predictions from VCC and ACP (CE

⁶We use NLTK POS tagger. Spacy POS tagger results in similar performance. More details in appendix.

⁷Note that we do not need additional image labels to train VCC, our labels are extracted automatically from the QA pairs. Same for ACP.

Table 8: Performance of GVQA (our model) compared to SAN on VQA-CP datasets. GVQA consistently outperforms SAN.

| Dataset | Model | Overall | Yes/No | Number | Other |
|-----------|-------------|--------------|--------------|--------------|--------------|
| VQA-CP v1 | GVQA (Ours) | 39.23 | 64.72 | 11.87 | 24.86 |
| | SAN [495] | 26.88 | 35.34 | 11.34 | 24.70 |
| VQA-CP v2 | GVQA (Ours) | 31.30 | 57.99 | 13.68 | 22.14 |
| | SAN [495] | 24.96 | 38.35 | 11.14 | 21.74 |

in the case of yes/no questions) on the training data. During training, we use ground truth labels for yes/no and non yes/no questions for the Question Classifier. During testing, we first run the Question Classifier to classify questions into yes/no and non yes/no. And feed the questions into their respective modules to obtain predictions on the test set. Please refer to appendix for implementation details.

5.2.3 Experiments on VQA-CP v1 and VQA-CP v2

Model accuracies: Table 5.2.3 shows the performance of our GVQA model in comparison to SAN (the model which GVQA is built off of) on VQA-CP v1 and VQA-CP v2 datasets using the VQA evaluation metric [27]. Accuracies are presented broken down into Yes/No, Number and Other categories. As it can be seen from Table 5.2.3, the proposed architectural improvements (in GVQA) over SAN show a significant boost in the overall performance for both the VQA-CP v1 (12.35%) and VQA-CP v2 (6.34%) datasets. It is worth noting that owing to the modular nature of the GVQA architecture, one may easily swap in other attention modules into the VCC. Interestingly, on the VQA-CP v1 dataset, GVQA also outperforms MCB [150] and NMN [24] (Tables 5.1.2 and 5.1.2) on the overall metric (mainly for yes/no questions), in spite of being built off of a relatively simpler attention module from SAN, and using relatively less powerful image features (VGG-16) as compared to ResNet-152 being used in MCB. On the VQA-CP v2 dataset, GVQA outperforms NMN in overall metric (as well as for number questions) and MCB for yes/no and number questions.

To check if our particular VQA-CP split was causing some irregularities in performance, we created four sets of VQA-CP v2 splits with different random seeds. This also led to a large portion of the dataset (84%) being covered across the test splits. The results show that GVQA consistently outperforms SAN across all four splits with average improvement being 7.14% (standard error: 1.36). Please see appendix for performance on each split.

Performance of Model Components *Question Classifier*: On the VQA-CP v1 test set, the LSTM based question classifier obtains 99.84% accuracy. *ACP*: The Top-1 test accuracy is 54.06%, with 84.25% for questions whose answers are in attribute clusters and 43.17% for questions whose answers are in object clusters. The Top-3 accuracy rises to 65.33%. Note that these accuracies are computed using the automatically created clusters. *VCC*: The weighted mean test F1 score across all classifiers is 0.53. The individual concepts are weighted as per the number of positive samples, reflecting the coverage of that concept in the test set. Please refer to appendix for accuracies on the VQA-CP v2 dataset.

5.2.4 Role of GVQA Components

In order to evaluate the role of various GVQA components, we report the experimental results (on VQA-CP v1) by replacing each component in GVQA (denoted by “<component>”) with its traditional counterpart, i.e., modules used in traditional VQA models (denoted by “+ <traditional counterpart>”). For instance, GVQA - CE + LSTM represents a model where CE in GVQA has been replaced with an LSTM. The results are presented in Table 5.2.4 along with the result of the full GVQA model for reference.

GVQA - Q_{main} + Q_{full} : GVQA’s performance when the entire question (Q_{full}) is fed into VCC (as opposed to after removing the question type (Q_{main})) is 33.55% (overall), which is 5.68% (absolute) less than that with Q_{main} . Note that even with

feeding the entire question, GVQA outperforms SAN, thus demonstrating that removing question type information helps but isn't the main factor behind the better performance of GVQA. As an additional check, we trained a version of SAN where the input is Q_{main} instead of Q_{full} . Results on VQA-CP v2 show that this version of SAN performs 1.36% better than the original SAN, however still 4.98% worse than GVQA (with Q_{main}). Please see appendix for detailed performance of this version of SAN.

GVQA - CE + LSTM: We replace CE with an LSTM (which is trained end-to-end with the Visual Verifier (VV) using VQA loss). The overall performance drops by 11.95%, with a drop of 28.76% for yes/no questions. This is an expected result, given that Table 5.2.3 shows that GVQA significantly outperforms SAN on yes/no questions and the CE is a crucial component of the yes/no pipeline.

GVQA - ACP + LSTM: We replace ACP with an LSTM (which is trained end-to-end with the Answer Predictor (AP) using VQA loss). The overall performance is similar to GVQA. But, the presence of ACP makes GVQA transparent and interpretable (see Section 5.2.6).

GVQA - VCC_{loss} : We remove the VCC loss and treat the output layer of VCC as an intermediate layer whose activations are passed to the Answer Predictor (AP) and trained end-to-end with AP using VQA loss. The overall performance improves by 1.72% with biggest improvement in the performance on other questions (3.19%). This suggests that introducing the visual concept (semantic) loss in between the model pipeline hurts. Although removing VCC loss and training end-to-end with VQA loss achieves better performance, the model is no longer transparent (see Section 5.2.6). Using VCC loss or not is a design choice one would make based on the desired accuracy vs. interpretability trade off.

GVQA - VCC_{loss} - ACP + LSTM: Replacing ACP with an LSTM on top of **GVQA - VCC_{loss}** hurts the overall performance by 2.09% with biggest drop

Table 9: Experimental results when each component in GVQA (denoted by “-<component>”) is replaced with its corresponding traditional counterpart (denoted by “+ <traditional counterpart>”).

| Model | Overall | Yes/No | Number | Other |
|----------------------------------|---------|--------|--------|-------|
| GVQA - Q_{main} + Q_{full} | 33.55 | 51.64 | 11.51 | 24.43 |
| GVQA - CE + LSTM | 27.28 | 35.96 | 11.88 | 24.85 |
| GVQA - ACP + LSTM | 39.40 | 64.72 | 11.73 | 25.33 |
| GVQA - VCC_{loss} | 40.95 | 65.50 | 12.32 | 28.05 |
| GVQA - VCC_{loss} - ACP + LSTM | 38.86 | 65.73 | 11.58 | 23.11 |
| GVQA | 39.23 | 64.72 | 11.87 | 24.86 |

Table 10: Results of GVQA and SAN on VQA v1 and VQA v2 when trained on the corresponding train splits.

| Model | VQA v1 | VQA v2 |
|----------------------|--------|--------|
| Oracle (GVQA, SAN) | 63.77 | 61.96 |
| Oracle (SAN, SAN) | 60.85 | 56.68 |
| Ensemble (GVQA, SAN) | 56.91 | 52.96 |
| Ensemble (SAN, SAN) | 56.56 | 52.45 |
| SAN | 55.86 | 52.02 |
| GVQA | 51.12 | 48.24 |

(4.94%) for “other” questions (see **GVQA - VCC_{loss}** and **GVQA - VCC_{loss} - ACP + LSTM** rows in Table 5.2.4). This suggests that ACP helps significantly (as compared to an LSTM) in the absence of VCC loss (and it performs similar to an LSTM in the presence of VCC loss, as seen above). In addition, ACP adds interpretability to GVQA.

5.2.5 Experiments on VQA v1 and VQA v2

We also trained and evaluated GVQA on train and val splits of the VQA v1 [27] and VQA v2 [168] datasets (results in Table 5.2.5⁸). On VQA v1, GVQA achieves 51.12% overall accuracy, which is 4.74% (absolute) less than SAN. This gap is not surprising because VQA v1 has well-established heavy language priors that existing

⁸We present overall and yes/no accuracies only. Please refer to appendix for performance on number and other categories.

models (including SAN) can “memorize” from train set and exploit on the test set (since test set contains same priors as train set), whereas GVQA is designed not to. As vision improves, grounded models like GVQA may show improved performance over models that leverage priors from training data. Moreover, it is important to note that the gain (GVQA acc - SAN acc) on VQA-CP v1 (12.35% absolute) is much higher than the loss (SAN acc - GVQA acc) on VQA v1 (4.74% absolute).

On VQA v2, GVQA under performs SAN by 3.78% overall, which is less than SAN acc - GVQA acc on VQA v1. And it outperforms SAN by 3.14% for yes/no questions. This shows that when the priors are weaker (in VQA v2 compared to those in VQA v1), the gap between GVQA and SAN’s performance decreases. We also trained and evaluated GVQA- VCC_{loss} on both the VQA v1 and VQA v2 datasets and found that it performs worse than GVQA on VQA v1 and similar to GVQA on VQA v2. So in addition to interpretability, GVQA is overall better than GVQA- VCC_{loss} on these original VQA splits.

In order to check whether GVQA has strengths complementary to SAN, we computed the oracle of SAN’s and GVQA’s performance – **Oracle (GVQA, SAN)**, i.e., we pick the predictions of the model with higher accuracy for each test instance. As it can be seen from Table 5.2.5, the Oracle (GVQA, SAN)’s overall performance is 7.91% higher than that of SAN for VQA v1 (9.94% for VQA v2) suggesting that GVQA and SAN have complementary strengths. Also, note that Oracle (GVQA, SAN) is higher than Oracle (SAN, SAN) for both VQA v1 and VQA v2, suggesting that GVQA’s complementary strengths are more than that of another SAN model (with a different random initialization).

Inspired by this, we report the performance of the ensemble of GVQA and SAN **Ensemble (GVQA, SAN)** in Table 5.2.5, where the ensemble combines the outputs from the two models using product of confidences of each model. We can see that Ensemble (GVQA, SAN) outperforms Ensemble (SAN, SAN) by 0.35% overall for

VQA v1 (and by 0.51% for VQA v2). It is especially better for yes/no questions. We also found that the ensemble of GVQA- VCC_{loss} with SAN performs worse than Ensemble (SAN, SAN) for both the VQA datasets (refer to appendix for accuracies). Hence, GVQA is a better complement of SAN than GVQA- VCC_{loss} , in addition to being more transparent.

5.2.6 Transparency



| | | | | | | |
|----------------------------------|--|--|---|--|------------------|----------------|
| Question | What sport are they playing ? | | | Is the person smiling ? | | |
| Image |  | | |  | | |
| Q-classifier | non yes/no | | | yes/no | | |
| Top ACP Cluster Predictions | # 3 (0.9884) tennis frisbee baseball | #16 (0.0046) surfing skateboarding parasailing | #19 (0.0040) skiing snowboarding downhill | ACP is deactivated. CE is activated. Extracted concept: smiling | | |
| Top VCC (per cluster) Prediction | baseball (0.962) | skateboarding (0.001) | skiing (0.0009) | Top VCC predictions for the cluster containing 'smiling' | | |
| | baseball (0.962) | skateboarding (0.001) | skiing (0.0009) | smiling (0.555) | woman (0.417) | man (0.190) |
| | baseball | ✓ | | yes | ✓ | |

Figure 24: Qualitative examples from GVQA. **Left:** We show top three answer cluster predictions (along with random concepts from each cluster) by ACP. Corresponding to each cluster predicted by ACP, we show the top visual concept predicted by VCC. Given these ACP and VCC predictions, the Answer Predictor (AP) predicts the correct answer ‘baseball’. **Right:** Smiling is the concept extracted by the CE whose visual presence in VCC’s predictions is verified by the Visual Verifier, resulting in ‘yes’ as the final answer.

The architecture design of GVQA makes it more transparent than existing VQA models because it produces interpretable intermediate outputs (the outputs of VCC, ACP and the concept string extracted by the CE) unlike most existing VQA models.



Figure 25: Left: GVQA’s prediction (*‘green’*) can be explained as follows – ACP predicts that the answer should be a *color*. Of the various visual concepts predicted by VCC, the only concept that is about color is *green*. Hence, GVQA’s output is *‘green’*. SAN incorrectly predicts *‘yellow’*. SAN’s architecture doesn’t facilitate producing an explanation of why it predicted what it predicted, unlike GVQA. Right: Both GVQA and SAN incorrectly answer the question. GVQA is incorrect perhaps because VCC predicts *‘black’*, instead of *‘gray’*. In order to dig further into why VCC’s prediction is incorrect, we can look at the attention map (in appendix), which shows that the attention is on the pants for the person’s left leg, but on the socks (black in color) for the person’s right leg. So, perhaps, VCC’s “black” prediction is based on the attention on the person’s right leg.

We show some example predictions from GVQA in Fig. 24. We can see that the intermediate outputs provide insights into why GVQA is predicting what it is predicting and hence enable a system designer to identify the causes of error. This is not easy to do in existing VQA models. Fig. 25 shows two other examples (one success and one failure) comparing and contrasting how GVQA’s intermediate outputs can help explain successes and failures (and thus, enabling targeted improvements) which is not possible to do for SAN and most other existing VQA models. See appendix for more such examples.

5.2.7 Conclusion

In this work, we proposed a novel Grounded Visual Question Answering model (GVQA) that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘cheating’ by primarily relying on priors in the training data. Specifically, GVQA explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question, enabling the model to more robustly generalize across different distributions of answers. GVQA is built off an existing VQA model – Stacked Attention Networks (SAN). Our experiments demonstrate that GVQA significantly outperforms SAN on both VQA-CP v1 and VQA-CP v2 datasets. Interestingly, it also outperforms more powerful VQA models such as Multimodal Compact Bilinear Pooling (MCB) in several cases. GVQA offers strengths complementary to SAN when trained and evaluated on the original VQA v1 and VQA v2 datasets. Finally, GVQA is more transparent and interpretable than existing VQA models.

GVQA is a first step towards building models which are visually grounded by design. Future work involves developing models that can utilize the best of both worlds (visual grounding and priors), such as, answering a question based on the knowledge about the priors of the world (sky is usually blue, grass is usually green) when the model’s confidence in the answer predicted as result of visual grounding is low.

5.3 *Adversarial Regularization for Visual Question Answering*

5.3.1 Introduction

In Section 5.2, we saw that GVQA is more robust to changing priors than existing VQA models. Although GVQA can be built on top of any existing VQA model, it does require non-trivial changes in the architecture. In this section, we propose

a simple drop-in regularizer for achieving robustness against changing priors. This regularizer can be added to any existing VQA model’s objective function, without requiring significant changes in the underlying VQA model’s architecture. Below we discuss the motivation and the intuition behind the proposed regularization scheme.

One intuitive measure of the strength of language priors in VQA is the performance of a ‘blind’ model that produces answers given only the question and not the associated image. In fact, this question-only model has become a standard and powerful baseline presented alongside VQA datasets [27, 167, 106, 218]. In this work, we codify this intuition, introducing a novel regularization scheme that sets a base VQA model against a question-only adversary to reduce the impact of language biases.

We consider unwanted language bias in VQA to be overly-specific relationships between questions and their likely answers learned from the training dataset – *i.e.* those that could enable a question-only model to achieve relatively high performance without ever seeing an image – and we explicitly optimize the question representation within a base VQA model to be uninformative to a question-only adversary model. In this adversarial regime, the question-only model is trained to answer as accurately as possible given the question encoding provided by the base VQA model; and simultaneously, the base VQA model is trained to adjust its question encoder (often implemented as a recurrent language model) to minimize the performance of the question-only model while maintaining its own VQA accuracy. Moreover, we leverage the question-only model to provide a differentiable notion of image grounding – the change in model confidence after considering the image – which we maximize explicitly for the VQA model. Thus, our objective consists of a question-only adversarial term and a difference of entropies term.

Our approach is largely model agnostic, end-to-end trainable, and simple to implement, consisting of a small, additional classification network built on the question representation of the base VQA model. We experiment on the VQA-CP dataset with

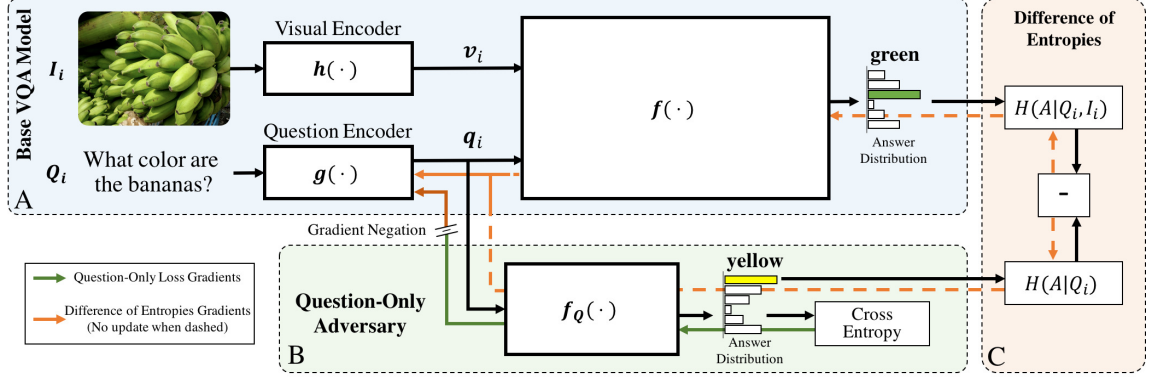


Figure 26: Given an arbitrary base VQA model (A), we introduce two regularizers. First, we build a question-only adversary (B) that takes the question embedding \mathbf{q}_i from the VQA model and is trained to output the correct answer from this information alone. For this network to succeed, \mathbf{q}_i must capture language biases from the dataset – the same biases that lead the base VQA model to ignore visual content. To reduce these biases, we set the base VQA model and the question-only adversary against each other, with the base VQA network modifying its question embedding to reduce question-only performance (shown here as gradient negation of the question-only model loss) Further, the question-only model allows estimation of the change in answer confidence given image (C), which we maximize explicitly.

multiple base VQA models, and find – 1) our approach provides consistent improvements over all baseline VQA models, 2) our approach outperforms the GVQA model significantly, 3) both question-only adversary and the difference of entropies components improve performance and their combination pushes this even further. On standard benchmarks [27, 167] where strong priors from training can be exploited on test set, our approach shows significantly smaller drops in accuracy compared to GVQA, with some settings facing only insignificant changes.

5.3.2 Reducing Language Bias Through Adversarial Regularization

Setting aside architectural specifics, the vast majority of VQA models operate on a set of similar design principles – first producing vector representations for the image and question and then combining them to predict the answer (often through complex attention mechanisms). However, when language biases are quite strong, the question feature may already be sufficiently discriminative and the model can learn to ignore

the visual signal without facing significant losses during training (*e.g.* “What color is the sky?” always mapping to “blue”). Such a model which fails to ground its answers in the image might be passable for benchmark datasets that carry similar biases; however, in the real-world, where brown grass and gray skies abound, its usefulness would be severely limited. In this section, we address this problem by explicitly reducing the discriminative power of the question feature – introducing a pair of adversarial regularizers that penalize the ability of a separate adversary network to confidently predict the answer from the question encoding alone.

Preliminaries. Given a dataset $\mathcal{D} = \{I_i, Q_i, a_i\}_{i=1}^N$ consisting of triplets of images $I_i \in \mathcal{I}$, questions $Q_i \in \mathcal{Q}$ and answers $a_i \in \mathcal{A}$, the VQA task is to learn a mapping $F: \mathcal{Q} \times \mathcal{I} \rightarrow [0, 1]^{|\mathcal{A}|}$ which produces an accurate distribution over the answer space given an input question-image pair.

Without loss of generality, we consider differentiable mappings that can be decomposed as an operation f over question and image encodings $g: \mathcal{Q} \rightarrow \mathbb{R}^d$ and $h: \mathcal{I} \rightarrow \mathbb{R}^k$ (as shown in Figure 26A). We write the prediction for instance i for this class of models as

$$\begin{aligned} \mathbf{v}_i &= h(I_i), \quad \mathbf{q}_i = g(Q_i) \\ P(\mathcal{A} \mid Q_i, I_i) &= f(\mathbf{v}_i, \mathbf{q}_i) \end{aligned} \tag{1}$$

where we denote the image and question embeddings as \mathbf{v}_i and \mathbf{q}_i respectively.

Nearly all existing VQA models follow this pattern. The image encoder $h(\cdot)$ is typically a fixed CNN pretrained on either classification or detection and the question encoder $g(\cdot)$ is usually some form of word or character level RNN learned during training. Typically these models are trained with standard cross-entropy, optimizing parameters to minimize (2) over the ground truth data.

$$\mathcal{L}_{VQA}(f, g, h) = \mathbb{E}_{\mathcal{I}, \mathcal{Q}, \mathcal{A}} [-\log P_f(a_i | Q_i, I_i)] \approx -\frac{1}{N} \sum_{i=1}^N \log f(\mathbf{v}_i, \mathbf{q}_i)[a_i] \tag{2}$$

Question-Only Model. One intuitive measure of the power of language priors in

VQA is the ability of a model to make low-error answer predictions from the question alone – in fact, some form of this ‘blind’ model has been frequently presented alongside VQA datasets for exactly this purpose [27, 167, 106, 218]. We formalize this question-only model as a mapping f_Q . As above, we assume f_Q is differentiable and operates on learned question encodings such that f_Q makes predictions

$$P_{f_Q}(\mathcal{A} \mid Q_i) = f_Q(\mathbf{q}_i), \quad \mathbf{q}_i = g(Q_i). \quad (3)$$

We parameterize this model as a simple two-layer neural network but note that arbitrary choices can be made in this regard. As above, this model can be trained with cross-entropy, minimizing

$$\mathcal{L}_{QA}(f_Q, g) = \mathbb{E}_{\mathcal{Q}, \mathcal{A}} [-\log P_{f_Q}(a_i \mid Q_i)] \approx -\frac{1}{N} \sum_{i=1}^N \log f_Q(\mathbf{q}_i)[a_i]. \quad (4)$$

5.3.2.1 *Adversarial Regularization with a Question-Only Adversary*

For any model of the form presented in (1), we can now introduce a simple adversarial regularizer that explicitly reduces the effect of language biases by modifying the question encoder to minimize the performance of this question-only adversary. Specifically, given a VQA model decomposed as f, g, h , we splice on the question-only model f_Q such that f_Q takes as input the encodings produced by $g(\cdot)$ (as in Figure 26), and establish opposing losses for the two networks which we detail below.

Learning the Question-Only Adversary. The question-only model f_Q is trained to minimize the cross-entropy loss \mathcal{L}_Q in (4); however, parameters in $g(\cdot)$ are not updated with respect to this loss – in effect, this forces f_Q to perform as well as possible given the question encodings produced by the question encoder $g(\cdot)$ from the base VQA model.

Adversarial Regularization for VQA. As performance of the question-only model acts as a proxy for the language biases represented in the question encodings $\mathbf{q}_i =$

$g(Q_i)$, one approach to reduce bias representation is to adjust $g(\cdot)$ such that the question-only model does poorly. As such, we can write this adversarial relationship between the question-only (f_Q) and base VQA models (f, g, h) as

$$\min_{f,g,h} \max_{f_Q} \mathcal{L}_{VQA}(f, g, h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g) \quad (5)$$

We note that in practice, training with this adversarial regularizer can be realized with a simple gradient negation of the question-only adversary’s loss as shown in Figure 26. Specifically, we back-propagate the negative of the gradient of $\mathcal{L}_Q(f_Q, g)$ accumulated at \mathbf{q}_i through the question encoder – updating the question encoder in a way that maximizes $\mathcal{L}_Q(f_Q, g)$.

The regularization coefficient $\lambda_Q \geq 0$ in (5) controls the trade-off between VQA performance and language bias reduction. For low values of λ_Q , little regularization occurs and the base model continues to learn language priors. On the other hand, large values of λ_Q force the model to remove all discriminative language biases, resulting in poor VQA performance for both the base VQA model and the question-only adversary – essentially stripping the question encoding of even basic question-type information (*e.g.* failing to learn that “*What color ... ?*” questions require color answers).

5.3.2.2 *An Adversarial Difference of Entropies Regularizer*

As the effect of this over-regularization for high-values of λ_Q highlights, the question-only adversary does not capture the full nuance of language bias in VQA. Given the question “*What color is the sky?*” it is reasonable to have a prior that the answer may be “*blue*”, but critically this belief should update depending on observations – *i.e.* the answer distribution should sharpen after viewing the image.

To capture this intuition, we add an additional term that aims to maximize the information gained about the answer from looking at the image. Specifically, we introduce another adversarial regularizer corresponding to the difference in entropies

between the base model prediction given the image and the question-only model which we write as

$$\mathcal{L}_H(f, g, h, f_Q) = \mathbb{E}_{I, Q} [H(\mathcal{A} \mid \mathcal{Q}) - H(\mathcal{A} \mid \mathcal{I}, \mathcal{Q})] \quad (6)$$

$$= \mathbb{E}_{q \sim P(\mathcal{Q})} [H(\mathcal{A} \mid q)] - \mathbb{E}_{q, v \sim P(\mathcal{Q}, \mathcal{I})} [H(\mathcal{A} \mid q, v)] \quad (7)$$

$$\approx \frac{1}{N} \sum_{i=1}^N (H(f_Q(\mathbf{q}_i)) - H(f(\mathbf{v}_i, \mathbf{q}_i))) \quad (8)$$

We note that this regularizer resembles the conditional mutual information (CMI) between the answer and image given the question $I(A; I|Q)$; however, $f_Q(q)$ is not constrained to be the marginal of $f(v, q)$ such that estimating the CMI in this way is ill-defined.

We can then update the adversarial relationship between f and f_Q from (5) with \mathcal{L}_{MI} , writing

$$\min_{f, g, h} \max_{f_Q} L_{VQA}(f, g, h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g) - \lambda_H \mathcal{L}_H(f, g, h, f_Q) \quad (9)$$

where $\lambda_H \geq 0$ controls the strength of the difference of entropies regularizer. Note that while \mathcal{L}_H is a function of f, g, h , and f_Q , we only update the parameters of the question encoding g based on this loss. Otherwise, f_Q could learn to produce sharp output distributions from arbitrary question features to minimize \mathcal{L}_H . Likewise, f or h can easily adjust to produce arbitrarily peaky outputs, which we observe can lead to significant over-fitting.

As before, the question-only adversary f_Q in this setting must still perform as well as possible given the question embedding from $g(\cdot)$, but this embedding is now additionally adjusted to maximize the entropy of f_Q 's output, while minimizing that of the VQA model. In the experiments that follow, we show that both of these adversarial regularizers improve performance on a language bias sensitive task. Further, we note that their benefits compound, with models combining both terms performing better across a wider range of regularization coefficients.

5.3.3 Experiments

Implementation. Our question-only adversary model is implemented as a 2-layer multi-layer perceptron with 256 hidden units and a ReLU activation that takes as input the question encoding from a base VQA network. The network’s output is a distribution over the candidate answers. We train the entire system (base VQA and question-only model) end-to-end with parameters initialized from scratch. We set batch size to 150, learning rate to 0.001, weight decay of 0.999 and use the Adam optimizer. The model takes ~ 8 hours to train on a TITAN X for SAN (Torch, ~ 60 epochs) and < 1 hour for UpDown (PyTorch, ~ 40 epochs). We use public codebases for both.

As discussed in Section 5.3.2, we update the parameters of the question encoding with respect to the VQA loss, the difference of entropies loss, and the negative of the question-only loss. The remaining VQA model parameters are trained with just the VQA loss. The question-only model is updated only by its VQA loss cross entropy loss term despite contributing to the difference of entropies loss.

Models. We evaluate the effect of our proposed regularization on the following base models:

- **Stacked Attention Network (SAN)** [493] – SAN encodes questions with a long short-term memory (LSTM) encoder and the image is encoded with a pretrained VGGNet [411]. The model performs two-hop question-based image attention and the final joint feature is passed to a 1000-way answer classifier. This model is trained with standard cross-entropy.
- **Bottom-Up and Top-Down Attention (UpDn)** [21] – Up-Down encodes questions with a gated recurrent unit (GRU) encoder and represents images as a set of bounding box features extracted from Faster R-CNN [371]. Soft-attention

over these regions is computed based on the question features and the attention-pooled feature is combined with the question as input to the classification layer. This model is trained directly on VQA score under a multi-label binary cross-entropy loss (see [21] for more details). We also apply this loss for the question-only model in our experiments, but compute a softmax over these outputs when computing entropies.

For both SAN⁹ and Up-Down¹⁰, we build on top of publicly available reimplementations. In the following results, we indicate the addition of our question-only adversarial regularization with **Q-Adv** and the difference of entropies term as **DoE**.

We also compare to the **GVQA** model built atop **SAN**. As we saw in Section 5.2, GVQA explicitly separates perception from question answering by introducing a Visual Concept Classifier (VCC) and an Answer Cluster Predictor (ACP). By design, this model isolates the answering module from the input question, mitigating the effect of language biases, but at a cost of relatively low standard VQA performance and multi-stage training.

Datasets and Evaluation. We train our models on the VQA-CP [13] train split and evaluate on the test set using the standard VQA evaluation metric [27]. For each model, we also report results when trained and evaluated on the standard VQA train and validation splits [27, 167] with the same regularization coefficients used for VQA-CP to compare with [13].

VQA-CP does not have a validation set and generating such a split is complicated by the need for it to contain priors different from both the training and test sets in order to be an accurate estimate of generalization under changing priors – an ill-defined notion for binary questions. As such, we set initial regularizer coefficients such that gradients at the question encoding are roughly equal in magnitude for all

⁹SAN Codebase: <https://github.com/abhshkdz/neural-vqa-attention>

¹⁰Up-Down Codebase: <https://github.com/hengyuan-hu/bottom-up-attention-vqa>

Table 11: Performance on VQA-CP v2 **test** and VQA v2 **val**. We significantly improve the accuracy of base models and achieve state-of-the-art performance on the VQA-CP dataset.

| Model | λ_Q | λ_H | VQA-CP v2 test | | | | VQA v2 val | | | |
|--------------|-------------|-------------|-----------------------|--------------|--------------|--------------|-------------------|--------|--------|-------|
| | | | Overall | Yes/No | Number | Other | Overall | Yes/No | Number | Other |
| GVQA [13] | - | - | 31.30 | 57.99 | 13.68 | 22.14 | 48.24 | 72.03 | 31.17 | 34.65 |
| SAN [493] | - | - | 24.96 | 38.35 | 11.14 | 21.74 | 52.41 | 70.06 | 39.28 | 47.84 |
| Ours SAN | 0.15 | - | 27.24 | 54.50 | 14.91 | 16.33 | 52.18 | 69.81 | 39.21 | 47.52 |
| | - | 25 | 25.75 | 42.21 | 12.08 | 20.87 | 52.38 | 70.05 | 39.64 | 47.41 |
| | 0.15 | 25 | 33.29 | 56.65 | 15.22 | 26.02 | 52.31 | 69.98 | 39.33 | 47.63 |
| UpDn [21] | - | - | 39.74 | 42.27 | 11.93 | 46.05 | 63.48 | 81.18 | 42.14 | 55.66 |
| Ours UpDn | 0.005 | - | 40.08 | 42.34 | 13.02 | 46.33 | 60.53 | 77.70 | 41.00 | 52.65 |
| | - | 0.05 | 40.43 | 42.62 | 12.19 | 47.03 | 63.43 | 81.15 | 42.64 | 55.45 |
| | 0.005 | 0.05 | 41.17 | 65.49 | 15.48 | 35.48 | 62.75 | 79.84 | 42.35 | 55.16 |

loss terms at the beginning of training and then explore a small region around this point. We report the best performing coefficients alongside our results and provide further analysis of the effect of these parameters in Section 5.3.4. Notably, we find these coefficients to be highly model dependent but generalize well between datasets and regularizer ablations. All models are trained until convergence as we have no validation set on which to base early-stopping.

5.3.4 Results

Table 11 presents our primary results on both the VQA-CP v2 and the VQA v2 datasets. Table 12 also shows limited results on the much more biased VQA v1 dataset [27] and its CP counterpart – VQA-CP v1 [13]. We make a number of observations below.

The proposed regularizers help, resulting in state-of-art performance on VQA-CP. For both SAN and UpDn models, adding the question-only adversary (Q-Adv) improves the performance of the respective base models (2.28% for SAN and 0.34% for UpDn) on the VQA-CP v2 dataset. Similarly, the difference of entropies (DoE) regularizer boosts the performance of both SAN and UpDn models, gaining improvements of 0.79% and 0.69% respectively. The combination of the Q-Adv and

Table 12: Performance on VQA-CP v1 **test** and VQA v1 **val**.

| Model | λ_Q | λ_H | VQA-CP v1 test | | | | VQA v1 val | | | |
|--|-------------|-------------|-----------------------|--------------|--------------|--------------|-------------------|--------|--------|-------|
| | | | Overall | Yes/No | Number | Other | Overall | Yes/No | Number | Other |
| GVQA [13] | - | - | 39.23 | 64.72 | 11.87 | 24.86 | 51.12 | 76.90 | 32.79 | 36.43 |
| SAN [493] | - | - | 26.88 | 35.34 | 11.34 | 24.70 | 55.86 | 78.54 | 33.46 | 44.51 |
| SAN + Q-Adv | 0.15 | - | 28.02 | 35.70 | 11.70 | 19.99 | 52.01 | 70.68 | 32.39 | 42.91 |
| Ours SAN + DoE | - | 25 | 27.83 | 36.33 | 11.15 | 24.03 | 54.08 | 78.19 | 32.59 | 41.44 |
| SAN + Q-Adv + DoE | 0.15 | 25 | 43.43 | 74.16 | 12.44 | 25.32 | 52.15 | 71.06 | 32.59 | 42.91 |

DoE regularizers further boosts the performance, resulting in 8.33% improvement over SAN and 1.43% over UpDn. Comparing our SAN + Q-Adv + DoE model to GVQA which is also built on top of SAN, we outperform GVQA significantly (1.99%). Our UpDn + Q-Only + DoE model also sets a new state-of-the-art on VQA-CP v2, improving over GVQA by 9.87% (although it is important to note the more powerful base architecture already outperforms GVQA by 8.44%).

Similar trends repeat for VQA-CP v1 as well. With the question-only regularizer improving SAN by 1.14%, DoE by 0.95%, and their combination by over 16.55% – outperforming GVQA by 4.2% and again setting state-of-the-art. We note that these larger gains are in part due to the increased language biases present in the VQA-CP v1 dataset.

Moreover, we find the question-only network performs increasingly poorly as our models perform better on VQA-CP – indicating that optimization is going well and that the intuition behind our regularizers seems well-founded.

The proposed regularizers do not hurt significantly on VQA v2. When trained and tested on the VQA v2 dataset (right side of Table 11), the addition of the proposed regularizers results in a insignificant drop in the performance for SAN (0.1%) and a minor drop in performance for UpDn (0.73%) compared to prior work. This is in contrast to GVQA, whose performance drops by 4.17% for SAN on VQA v2 (note that GVQA is built off of SAN).

The more the biases, the higher the gain on VQA-CP, and the higher the

loss on VQA. VQA v1 has significantly more bias than VQA v2 and consequentially VQA-CP v1 has a sharper change between training and test. As such, we observe the proposed regularizers improve over the base model significantly more in VQA-CP v1 (16.55% for SAN) than in VQA-CP v2 (8.33% for SAN). For the same reasons, the proposed regularizers hurt a bit more on VQA v1 (3.71% for SAN compared to 0.1% on VQA v2), where strong language biases can be leveraged to boost performance. However, this drop in the performance on VQA v1 is still significantly less than that with GVQA (4.74%).

UpDn [21] is less driven by biases than SAN. The drop in the performance of UpDn from VQA v2 to VQA-CP v2 is 23.74% which is significantly less than that of SAN (27.45%). This shows that UpDn may be less driven by biases than SAN. And hence, the gains in UpDn (1.43%) due to the proposed regularizers are less than those in SAN (8.33%).

Our approach results in less biased output distributions. Figure 28 shows answer frequency distributions for VQA v2 train, SAN, our SAN+Q-Adv+DoE model (marked Ours), and VQA v2 test for three questions: “*What color is the dress she/he is wearing?*”, “*What sport ...?*” “*What color is the fire hydrant?*”. It is quite clear that while neither of the SAN based models completely match the test distribution, the base SAN model aligns significantly more with the training distribution – even amplifying the bias for ‘blue’ in the first question despite very few answers being ‘blue’ in test.

Difference of entropies (DoE) stabilizes training with the question-only adversary. Figure 27 shows VQA-CP v2 test performance of the SAN model, for a range of question-only regularizer coefficients λ_Q . We can see that when the DoE term is not used (orange line), performance begins to drop after approximately 0.2 and by 0.35 has deteriorated significantly. At these higher values, nearly all discriminative information in the question encoding is lost – with the VQA model sacrificing its own

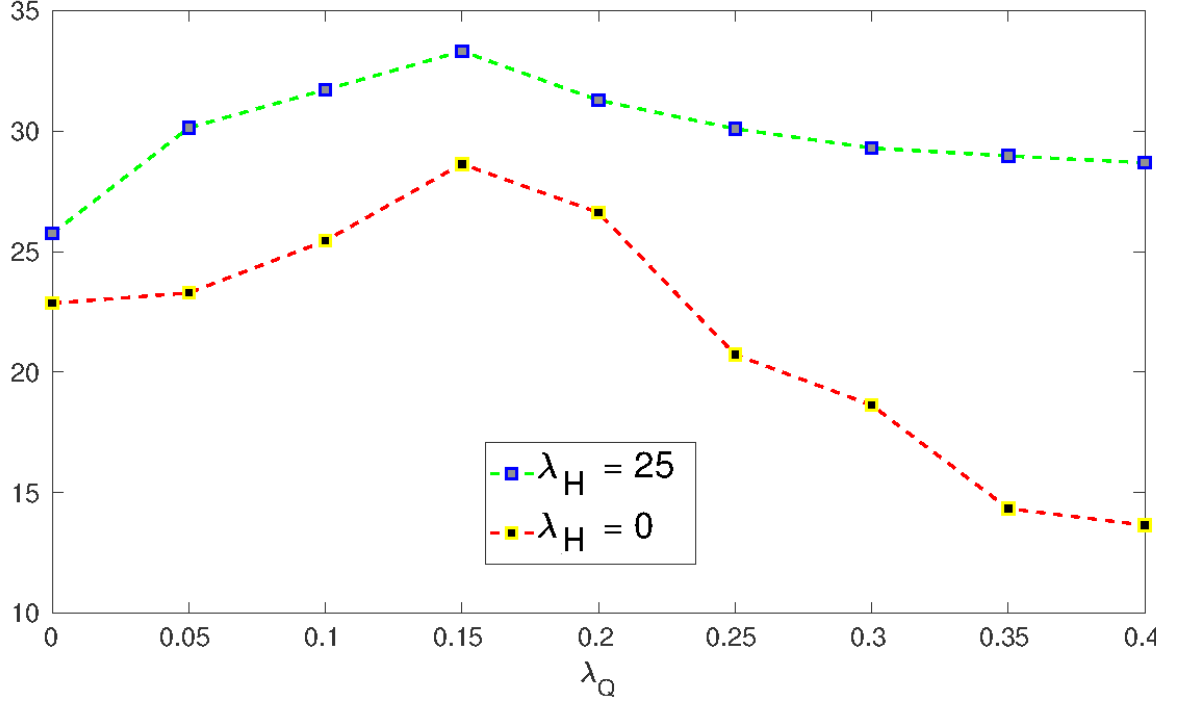


Figure 27: Maximizing difference of entropies (DoE) along with the question-only adversarial regularization for the SAN model, not only improves results on changing priors, but also stabilizes training.

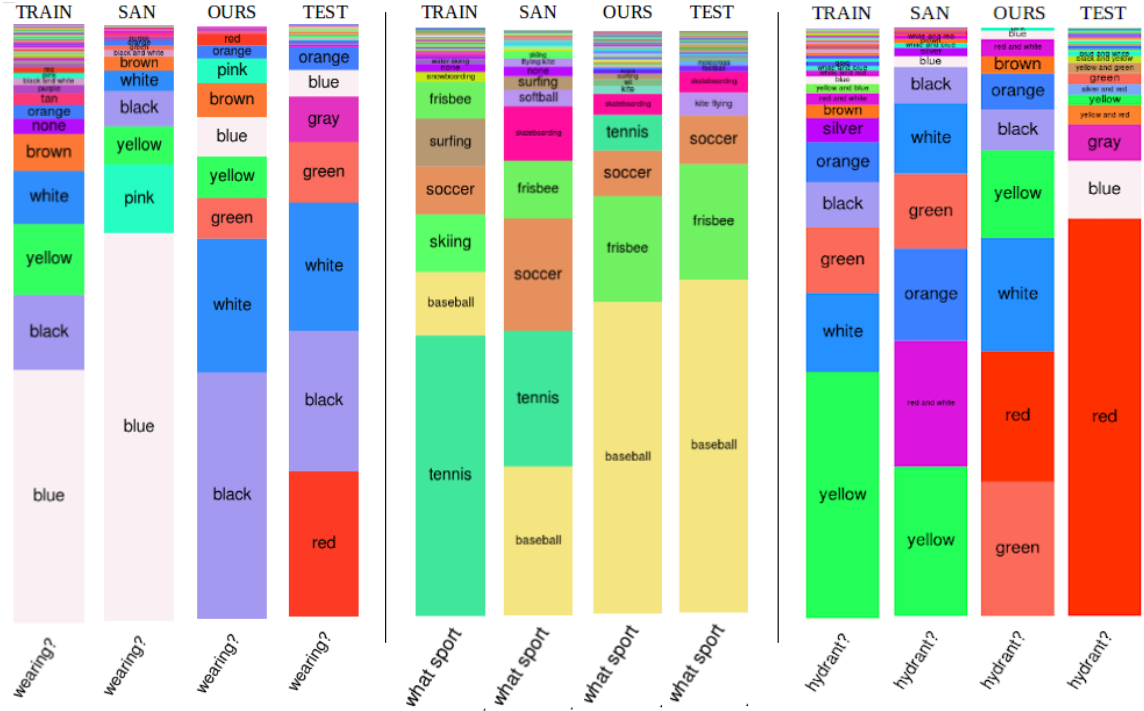


Figure 28: Answer distribution for SAN+Q-Adv+DoE mimic the prior less for questions with high language bias.

performance to lower that of the question-only model. However, we observe that for reasonable values of λ_H , the strength of the question-only adversary can be varied over a much wider range with less dramatic losses (blue curve in Figure 27). We observe a similar trend when keeping λ_Q constant and sweeping over λ_H , wherein a dramatic improvement is observed when moving to non-zero λ_H and then a slow decay for large values of λ_H . Unlike the question-only adversary, the DoE regularizer simultaneously seeks to sharpen the VQA models posterior while weakening the question-only prior.

Question-only performance: We study the performance of the question-only model after being trained on VQA-CP v2 using our regularizers. We compare to a question-only model trained without these regularizers, i.e. a model trained to predict the correct answer given the question-encoding learned by the base VQA model. We find this Q-only(SAN) model achieves 24.84% on the VQA-CP v2 training set compared to 13.85% for our SAN+Q-only+DoE model, demonstrating that our approach has effectively restricted the discriminative information in the question encoding.

Proposed model shows complementary strengths with the base model: To study whether our models learn complementary strengths to the base VQA models, we experiment with ensembles of both models. First, we consider oracle ensembles where the best model output for each data point is considered for evaluation. This is an upper bound on ensemble performance that relies on knowing ground truth. We find that the Oracle(Ours, SAN) ensemble outperforms two separately trained SAN models Oracle(SAN, SAN), by 1.48% for VQA v1 and by 3.46% for VQA v2—significantly lower gains than with Oracle(GVQA, SAN) which improves by 5.28%. It is notable however that the architecture of GVQA is significantly different from the base SAN model and hence is expected to exhibit different error patterns and a higher Oracle accuracy. To take a more attainable view, we also computed a standard ensemble Ensemble(Ours, SAN) and compared to an Ensemble(SAN, SAN) model,

outperforming it by 1.24% for VQA v2 but falling short by 0.15% for VQA v1. In contrast, Ensemble(GVQA, SAN) improves VQA v2 performance by only 0.54%.

5.3.5 Conclusion

We propose a novel adversarial regularization scheme for reducing the memorization of dataset biases in VQA based on a question-only adversary and the difference of model confidences after processing the image. Experiments on the VQA-CP dataset, show that this technique allows existing VQA models to significantly improve performance in the midst of changing priors. Consequently, we achieve state-of-the-art performance on VQA-CP. Our approach can be implemented as a simple, drop-in module on top of existing VQA models and easily trained end-to-end from scratch.

CHAPTER VI

CONCLUSION

In this dissertation, I introduce and study a multi-modal Artificial Intelligence (AI) task called Visual Question Answering (VQA) – given an image and a natural language question about the image (e.g., ‘*What kind of store is this?*’, ‘*Is it safe to cross the street?*’), the machine’s task is to automatically produce an accurate natural language answer (‘*bakery*’, ‘*yes*’). Specifically, my colleagues and I introduced the task of free-form and open-ended Visual Question Answering (VQA). We collected a large scale dataset (>0.25M images, >0.76M questions, ~10M answers) and made it publicly available (www.visualqa.org). This dataset, together with the development of baseline models and organization of annual challenges and workshops by us, led to remarkable improvements in the state-of-art on VQA. As part of my dissertation, I also developed novel techniques to characterize the behavior of VQA models. And finally, I addressed the issue of VQA models being driven by superficial correlations in training data and lacking sufficient image grounding by – 1) proposing a new evaluation protocol to evaluate the degree of visual groundedness in VQA models, 2) proposing a novel Grounded VQA (GVQA) model that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘cheating’ by primarily relying on priors in the training data, 3) proposing a novel adversarial regularization scheme that can be added to any existing VQA model’s objective function, without significantly changing the underlying VQA model’s architecture.

Future Work Directions: VQA and Beyond

VQA. Despite tremendous progress in VQA, there are some specific types of questions in VQA where the community has not made enough progress (mentioned below).

Some of these are highlighted below:

- **Counting:** The performance of state-of-art VQA models on counting questions (e.g., ‘*How many people are standing in the queue?*’, ‘*How many slices of pizza are there?*’) is only $\sim 45\%$ (compared to the human performance of $\sim 83\%$ and overall (across all questions) performance of state-of-art models of $\sim 71\%$). Clearly, the act of counting itself is not challenging – what is challenging is to parse the language, identify the referring expressions, grounding the referring expressions into visual concepts (e.g., detecting each individual slice of pizza, detecting each person who is standing and in the queue) – all of these are studied today as separate tasks in the computer vision and NLP community. I think studying unified VQA models that include these components as modules could be a step towards improving counting performance of VQA models.
- **Optical Character Recognition (OCR):** Another class of questions where the community has not made enough progress is all questions that require reading text photographed in images (e.g., ‘*What does the street sign say?*’, ‘*What is the name of the building?*’). Answering such questions requires the ability to do Optical Character Recognition (OCR), which I believe current VQA models lack because they do not get enough training signal from the downstream VQA loss to be able to learn OCR. Also, most of the existing VQA models predict answers by doing classification over a fixed set of K (typically 1000 - 3000) answers. Hence these models will not be able to correctly answer such OCR type of questions if the correct word / phrase does not lie in the list of those K answers. I believe building VQA models that use existing OCR techniques as modules and that have the ability to predict answers not seen during training could be a good first step towards making progress in this direction. Recently, with the release of the TextVQA dataset [412] and the organization of the TextVQA challenge (<https://textvqa.org/challenge>), the community

has already started making progress in this direction, but there is still a lot of room for improvement.

- **Knowledge Based Reasoning:** We have not made much progress on questions that require knowledge based reasoning and common sense (e.g., ‘*Does this person have 20/20 vision?*’, ‘*Is this food healthy?*’). Such questions require an agent to understand what ‘20/20’ vision means and what types of food items are healthy, in addition to understanding the visual content – recognizing that the person is wearing spectacles and that it is a fast food item. I believe a limiting factor in this area is lack of existence of a large-scale and open-ended Knowledge Based (KB) VQA dataset. Creating a much larger and open-ended KB-VQA dataset has the potential to push the progress in this direction.

Beyond: From Vision and Language to Actions. Most of my past work has been towards building agents that can ‘*see*’ and ‘*talk*’. However, for a lot of practical applications (e.g., physical agents navigating inside our houses executing natural language commands) we need agents that can not only ‘*see*’ and ‘*talk*’ but can also take actions. Below are some directions towards generalizing vision and language agents to be able to take actions.

In this space of building agents that can ‘*see*’, ‘*talk*’ and ‘*act*’, one bold initiative which was well ahead of its time was the SHRDLU [477] project, studied by Terry Winograd in 1972 – an agent that operates on a table top scene consisting of several blocks such as cuboid, cone, containers etc.; there is a teacher which instructs the agent what to do (e.g., ‘*pick up a red block*’) and the agent either executes an action or asks a question (if the instruction is not clear) (e.g., ‘*By “It”, I assume you mean the block which is taller than the one I am holding.*’’) (Fig. 29).

However, SHRDLU was a hand-engineered rule-based system. I think building a learning based SHRDLU agent can be a first step towards building agents that can ‘*see*’, ‘*talk*’ and ‘*act*’.

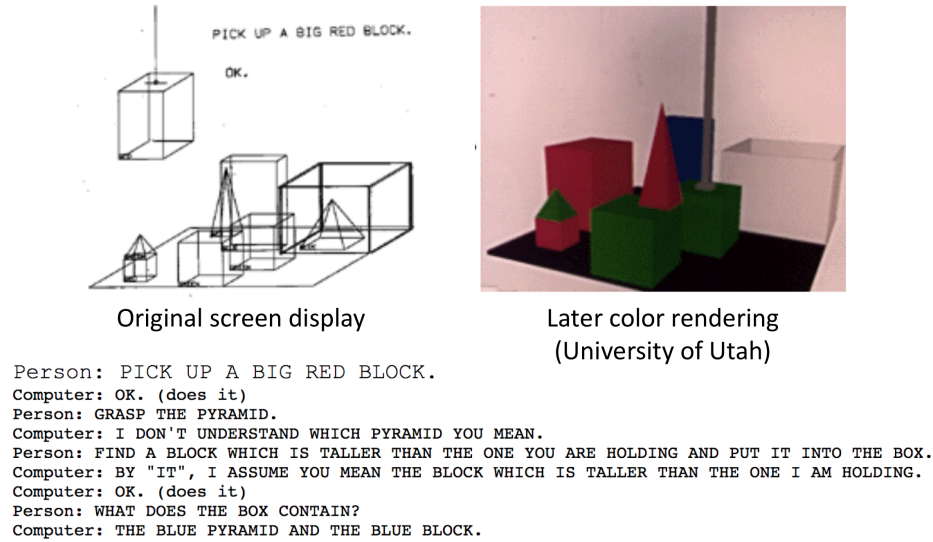


Figure 29: The table-top setup and an example dialog from the SHRDLU [477] project (studied by Terry Winograd in 1972).

As a baby step in this direction, in a recent work [19], I worked on the following - how can we train agents to follow language instructions grounded in visual data (e.g., ‘Add a red sphere’, ‘Add a large cylinder’) and execute actions to generate scenes that are consistent with the given instruction (Fig. 30). Using reinforced adversarial learning framework [154], we have taken the first step towards training agents that can follow simple instructions (as mentioned above).

More generally, in the long term, I look forward to interactive agents that can, for instance, edit images based on natural language queries such as ‘Change the background to winter.’, AI assistants such as Alexa that can not only process language commands but can also situate itself in its surrounding environment and can answer questions such as ‘Alexa, is my laptop in my bedroom?’, and finally agents that can move around our houses and execute natural language commands such as ‘Could you please get my laptop from upstairs?’.

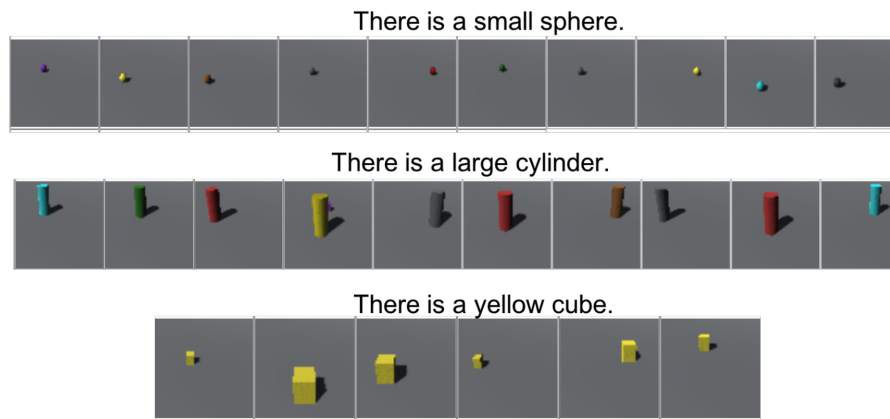


Figure 30: Given an instruction (*‘There is a small sphere’*), the the task for an agent is to execute actions to create scenes that are consistent with the given instruction (i.e., each such scene consists of a small sphere).

APPENDIX A

APPENDIX FOR VISUAL QUESTION ANSWERING

In this appendix, we provide:

1. - Additional analysis comparing captions and Q&A data
2. - Qualitative visualizations for “What is” questions
3. - Human accuracy on multiple-choice questions
4. - Details on VQA baselines
5. - “Age” and “Commonsense” of our model
6. - Details on the abstract scene dataset
7. - User interfaces used to collect the dataset
8. - List of the top answers in the dataset
9. - Additional examples from the VQA dataset

A.1 Captions vs. Questions

Do questions and answers provide further information about the visual world beyond that captured by captions? One method for determining whether the information captured by questions & answers is different from the information captured by captions is to measure some of the differences in the word distributions from the two datasets. We cast this comparison in terms of nouns, verbs, and adjectives by extracting all words from the caption data (MS COCO captions for real images and captions collected by us for abstract scenes) using the Stanford part-of-speech (POS)¹ tagger [437]. We normalize the word frequencies from captions, questions, and answers per image, and

¹Noun tags begin with NN, verb tags begin with VB, adjective tags begin with JJ, and prepositions are tagged as IN.

compare captions *vs.* questions and answers combined. Using a Kolmogorov-Smirnov test to determine whether the underlying distributions of the two datasets differ, we find a significant difference for all three parts of speech ($p < .001$) for both real images and abstract scenes. This helps motivate the VQA task as a way to learn information about visual scenes; although both captions and questions & answers provide information about the visual world, they do it from different perspectives, with different underlying biases [165], and can function as complementary to one another.

We illustrate the similarities and differences between the word distributions in captions *vs.* questions & answers as Venn-style word clouds [100] with size indicating the normalized count – Fig. 32 (nouns), Fig. 33 (verbs), and Fig. 34 (adjectives) for real images and Fig. 35 (nouns), Fig. 36 (verbs), and Fig. 37 (adjectives) for abstract scenes.² The left side shows the top words in questions & answers, the right the top words in captions, and the center the words common to both, with size indicating the harmonic mean of the counts.

We see that adjectives in captions capture some clearly visual properties discussed in previous work on vision to language [311], such as material and pattern, while the questions & answers have more adjectives that capture what is usual (*e.g.*, “dominant”, “approximate”, “higher”) and other kinds of commonsense properties (*e.g.*, “edible”, “possible”, “unsafe”, “acceptable”). Interestingly, we see that question & answer nouns capture information about “ethnicity” and “hairstyle”, while caption nouns capture information about pluralized visible objects (*e.g.*, “cellphones”, “daughters”) and groups (*e.g.*, “trio”, “some”), among other differences. “Man” and “people” are common in both captions and questions & answers.

One key piece to understanding the visual world is understanding spatial relationships, and so we additionally extract spatial prepositions and plot their proportions in the captions *vs.* the questions & answers data in Fig. 31 (left) for real images and

²Visualization created using <http://worditout.com/>.

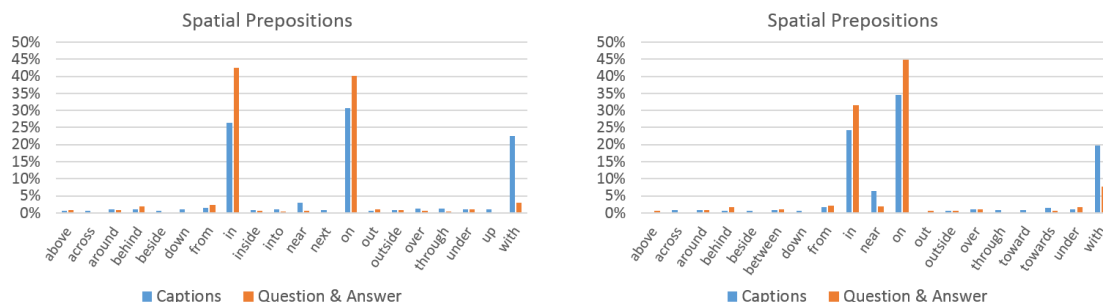


Figure 32: Venn-style word clouds [100] for nouns with size indicating the normalized count for real images.

Fig. 31 (right) for abstract scenes. We see that questions & answers have a higher proportion of specific spatial relations (*i.e.*, “in”, “on”) compared to captions, which have a higher proportion of general spatial relations (*i.e.*, “with”, “near”).

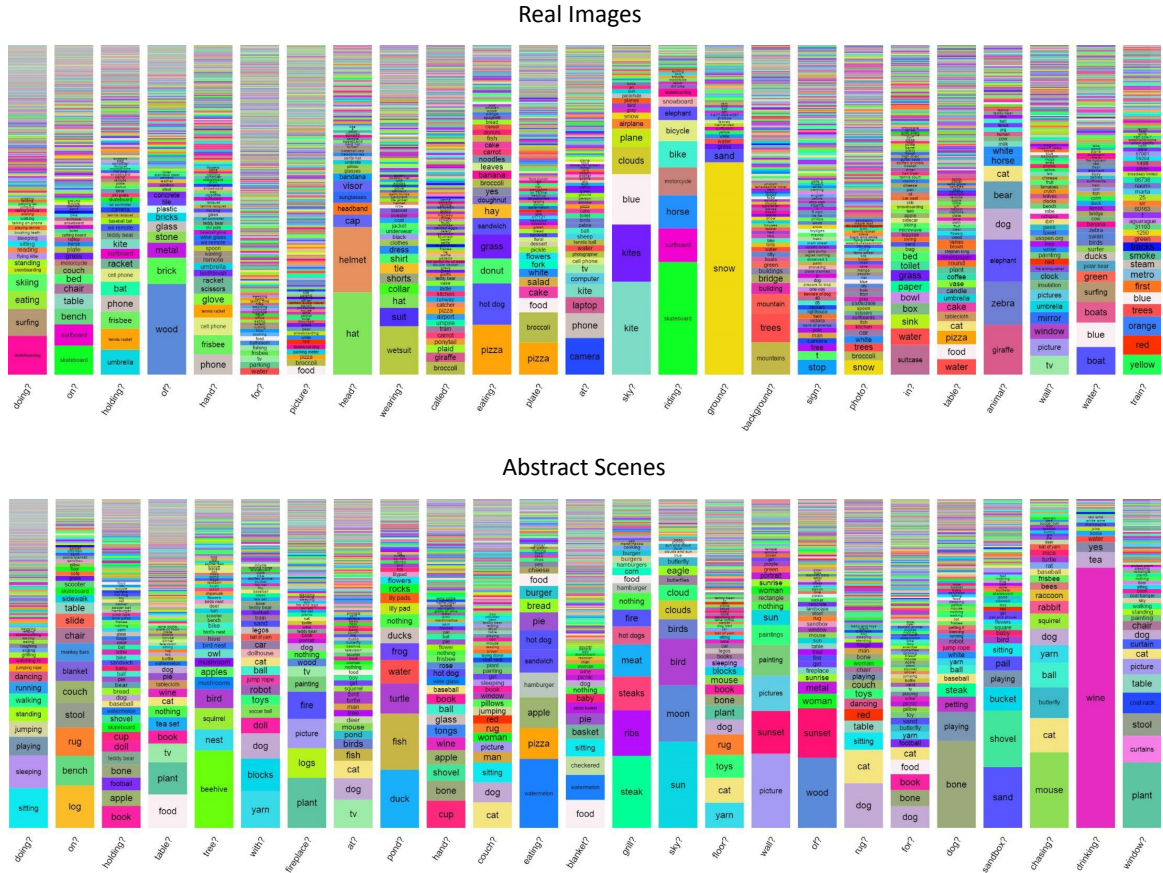


Figure 39: Distribution of answers for questions starting with “What is” for a random sample of 60K questions for real images (top) and all questions for abstract scenes (bottom). Each column corresponds to questions ending in different words, such as “doing?”, “on?”, *etc.*

Table 13: For each of the two datasets, real and abstract, first two rows are the human accuracies for multiple-choice questions when subjects were shown both the image and the question. Majority vote means we consider the answer picked by majority of the three subjects to be the predicted answer by humans and compute accuracy of that answer for each question. Average means we compute the accuracy of each of the answers picked by the subjects and record their average for each question. The last row is the inter-human agreement for open-ended answers task when subjects were shown both the image and the question. All accuracies are evaluated on a random subset of 3000 questions.

| Dataset | Accuracy Metric | All | Yes/No | Number | Other |
|----------|------------------|-------|--------|--------|-------|
| Real | MC majority vote | 91.54 | 97.40 | 86.97 | 87.91 |
| | MC average | 88.53 | 94.40 | 84.99 | 84.64 |
| | Open-Ended | 80.62 | 94.78 | 78.46 | 69.69 |
| Abstract | MC majority vote | 93.57 | 97.78 | 96.71 | 88.73 |
| | MC average | 90.40 | 94.59 | 94.36 | 85.32 |
| | Open-Ended | 85.66 | 95.32 | 94.17 | 74.12 |

questions. Table A.3 also shows the inter-human agreement for open-ended answer task. In comparison to open-ended answer, the multiple-choice accuracies are more or less same for “yes/no” questions and significantly better ($\approx 15\%$ increase for real images and $\approx 11\%$ increase for abstract scenes) for “other” questions. Since “other” questions may be ambiguous, the increase in accuracy using multiple choice is not surprising.

A.4 Details on VQA baselines

“per Q-type prior” baseline. We decide on different question types based on first few words of questions in the real images training set and ensure that each question type has at least 30 questions in the training dataset. The most popular answer for each question type is also computed on real images training set.

“nearest neighbor” baseline. For every question in the VQA test-standard set, we find its k nearest neighbor questions in the training set using cosine similarity in Skip-Thought [230] feature space. We also experimented with bag of words and

Word2Vec [303] feature spaces but we obtained the best performance with Skip-Thought. In this set of k questions and their associated images, we find the image which is most similar to the query image using cosine similarity in fc7 feature space. We use the fc7 features from the caffe net model in BVLC Caffe [207]. The most common ground truth answer of this most similar image and question pair is the predicted answer for the query image and question pair. We pick $k = 4$ on the test-dev set.

A.5 “Age” and “Commonsense” of our model

We estimate the age and degree of commonsense of our **best model** (deeper LSTM Q + norm I), selected using VQA test-dev accuracies). To estimate the age, we compute a weighted average of the average age per question, weighted by the accuracy of the model’s predicted answer for that question, on the subset of questions in the VQA validation set for which we have age annotations (how old a human needs to be to answer the question correctly). To estimate the degree of commonsense, we compute a weighted average of the average degree of commonsense per question, weighted by the accuracy of the model’s predicted answer for that question, on the subset of questions in the VQA validation set for which we have commonsense annotations (whether the question requires commonsense to answer it).

A.6 Abstract Scenes Dataset

In Fig. 40 (left), we show a subset of the objects that are present in the abstract scenes dataset. For more examples of the scenes generated, please see Fig. 45. The user interface used to create the scenes is shown in Fig. 40 (right). Subjects used a drag-and-drop interface to create the scenes. Each object could be flipped horizontally and scaled. The scale of the object determined the rendering order of the objects. Many objects have different attributes corresponding to different poses or types. Most animals have five different discrete poses. Humans have eight discrete expressions and

their poses may be continuously adjusted using a “paperdoll” model [29].

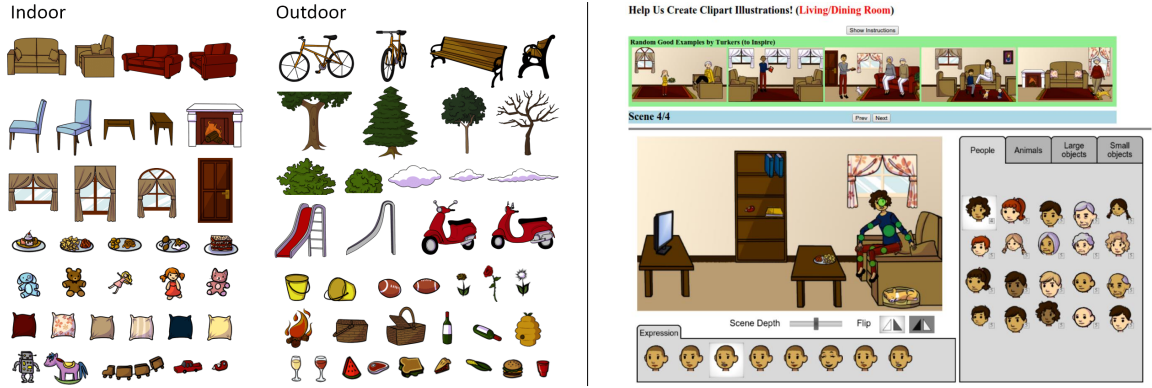


Figure 40: Left: A small subset of the objects present in the abstract scene dataset. Right: The AMT interface for collecting abstract scenes. The light green circles indicate where users can select to manipulate a person’s pose. Different objects may be added to the scene using the folders to the right.

A.7 User Interfaces

In Fig. 41, we show the AMT interface that we used to collect questions for images. Note that we tell the workers that the robot already knows the answer to the previously asked question(s), inspiring them to ask different kinds of questions, thereby increasing the diversity of our dataset.


Fig. 42 shows the AMT interface used for collecting answers to the previously collected questions when subjects were shown the corresponding images. Fig. 43 shows the interface that was used to collect answers to questions when subjects were not shown the corresponding image (*i.e.*, to help in gathering incorrect, but plausible, answers for the multiple-choice task and to assess how accurately the questions can be answered using common sense knowledge alone).

Stump a smart robot! Ask a question about this scene that a human can answer, but a smart robot probably can't!

Updated instructions: Please read carefully

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! **In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.**

Ask a question about this scene that this SMART robot probably can not answer, but any human can easily answer while looking at the scene in the image. **IMPORTANT:** The question should be about this scene. That is, the human should need the image to be able to answer the question – the human should not be able to answer the question without looking at the image.



Your work will get rejected if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed** below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please [ask about something different](#).
- Do not repeat questions.** Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a [new question each time](#) specific to the scene in each image.
- Each question should be a **single question**. **Do not ask questions that have multiple parts** or multiple sub-questions in them.
- Do not ask generic questions** that can be asked of many other scenes. Ask questions [specific to the scene in each image](#).

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer "if" looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3:


Page 2/3

Figure 41: Our AMT interface for collecting the third question for an image, when subjects were shown previous questions that were collected and were asked to ask a question different from previous questions.

Help Us Answer Questions About Images!
Updated instructions: Please read carefully

Please answer some questions about images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

If you don't follow the following instructions, your work will be rejected.



Your work will get rejected if you do not follow the instructions below:

- Answer the question based on what is going on in **the scene depicted in the image**.
- Your answer should be a **brief phrase** (not a complete sentence).
 - "It is a kitchen." -> "kitchen"
- For yes/no questions, please **just say yes/no**.
 - "You bet it is!" -> "yes"
- For numerical answers, please use **digits**.
 - "Ten." -> "10"
- If you need to speculate (e.g., "What just happened?"), provide an answer **that most people would agree on**.
- If you don't know the answer (e.g., specific dog breed), provide **your best guess**.
- Respond matter-of-factly and **avoid using conversational language or inserting your opinion**.

Please answer the question using as few words as possible:

Q1: What is unusual about this mustache?

A1:

Do you think you were able to answer the question correctly?
(Clicking an option will take you to the next question.)

Page 1/2

Figure 42: The AMT interface used to collect answers to a question when subjects were shown the image while answering the question.

Help Us Answer Questions!
Updated instructions: please read carefully

We will show you a series of questions **about possibly different scenes**. Your task is to answer them. Here's the catch: we will not show you the scenes!

So how can you answer the question correctly? Well, you can't. But your job is to **provide a plausible answer to the question**. What this means is the following: If we show the question alongside your answer to someone else (who also can't see the scene), they should think your answer "could be" correct.

Please keep your answer **brief**. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

If you don't follow the following instructions, your work will be rejected.

Instructions:

- Your answer should be a **brief phrase** (not a complete sentence).
 - "It is a kitchen." -> "kitchen"
- For yes/no questions, please **just say yes/no**.
 - "You bet it is!" -> "yes"
- For numerical answers, please use **digits**.
 - "Ten." -> "10"
- Respond matter-of-factly and **avoid using conversational language**.

Please provide a plausible answer to the question using as few words as possible:

Q1: What is unusual about this mustache?

A1:

How likely do you think it is that someone else would answer this question with the same answer as yours?
(Clicking an option will take you to the next question.)

Page 1/2

Figure 43: The AMT interface used to collect answers to a question when subjects were not shown the image while answering the question using only commonsense to collect the plausible, but incorrect, multiple-choice answers.

A.8 Answer Distribution

The top 250 answers in our real images dataset along with their counts and percentage counts are given below. The answers have been presented in different colors to show the different Part-of-Speech (POS) tagging of the answers with the following color code: **yes/no**, **noun**, **verb**, **adjective**, **adverb**, and **numeral**.

“yes” (566613, 22.82%), “no” (381307, 15.35%), “2” (80031, 3.22%), “1” (46537, 1.87%), “white” (41753, 1.68%), “3” (41334, 1.66%), “red” (33834, 1.36%), “blue” (28881, 1.16%), “4” (27174, 1.09%), “green” (22453, 0.9%), “black” (21852, 0.88%), “yellow” (17312, 0.7%), “brown” (14488, 0.58%), “5” (14373, 0.58%), “tennis” (10941, 0.44%), “baseball” (10299, 0.41%), “6” (10103, 0.41%), “orange” (9136, 0.37%), “0” (8812, 0.35%), “bathroom” (8473, 0.34%), “wood” (8219, 0.33%), “right” (8209, 0.33%), “left” (8058, 0.32%), “frisbee” (7671, 0.31%), “pink” (7519, 0.3%), “gray” (7385, 0.3%), “pizza” (6892, 0.28%), “7” (6005, 0.24%), “kitchen” (5926, 0.24%), “8” (5592, 0.23%), “cat” (5514, 0.22%), “skiing” (5189, 0.21%), “skateboarding” (5122, 0.21%), “dog” (5092, 0.21%), “snow” (4867, 0.2%), “black and white” (4852, 0.2%), “skateboard” (4697, 0.19%), “surfing” (4544, 0.18%), “water” (4513, 0.18%), “giraffe” (4027, 0.16%), “grass” (3979, 0.16%), “surfboard” (3934, 0.16%), “wii” (3898, 0.16%), “kite” (3852, 0.16%), “10” (3756, 0.15%), “purple” (3722, 0.15%), “elephant” (3646, 0.15%), “broccoli” (3604, 0.15%), “man” (3590, 0.14%), “winter” (3490, 0.14%), “stop” (3413, 0.14%), “train” (3226, 0.13%), “9” (3217, 0.13%), “apple” (3189, 0.13%), “silver” (3186, 0.13%), “horse” (3159, 0.13%), “banana” (3151, 0.13%), “umbrella” (3139, 0.13%), “eating” (3117, 0.13%), “sheep” (2927, 0.12%), “bear” (2803, 0.11%), “phone” (2772, 0.11%), “12” (2633, 0.11%), “motorcycle” (2608, 0.11%), “cake” (2602, 0.1%), “wine” (2574, 0.1%), “beach” (2536, 0.1%), “soccer” (2504, 0.1%), “sunny” (2475, 0.1%), “zebra” (2403, 0.1%), “tan” (2402, 0.1%), “brick” (2395, 0.1%), “female” (2372, 0.1%), “bananas” (2350, 0.09%), “table” (2331, 0.09%), “laptop” (2316, 0.09%), “hat” (2277, 0.09%), “bench” (2259,

0.09%), “flowers” (2219, 0.09%), “woman” (2197, 0.09%), “male” (2170, 0.09%),
 “cow” (2084, 0.08%), “food” (2083, 0.08%), “living room” (2022, 0.08%), “bus”
 (2011, 0.08%), “snowboarding” (1990, 0.08%), “kites” (1979, 0.08%), “cell phone”
 (1943, 0.08%), “helmet” (1885, 0.08%), “maybe” (1853, 0.07%), “outside” (1846,
 0.07%), “hot dog” (1809, 0.07%), “night” (1805, 0.07%), “trees” (1785, 0.07%),
 “11” (1753, 0.07%), “bird” (1739, 0.07%), “down” (1732, 0.07%), “bed” (1587,
 0.06%), “camera” (1560, 0.06%), “tree” (1547, 0.06%), “christmas” (1544, 0.06%),
 “fence” (1543, 0.06%), “nothing” (1538, 0.06%), “unknown” (1532, 0.06%), “tennis
 racket” (1525, 0.06%), “red and white” (1518, 0.06%), “bedroom” (1500, 0.06%),
 “bat” (1494, 0.06%), “glasses” (1491, 0.06%), “tile” (1487, 0.06%), “metal” (1470,
 0.06%), “blue and white” (1440, 0.06%), “fork” (1439, 0.06%), “plane” (1439, 0.06%),
 “airport” (1422, 0.06%), “cloudy” (1413, 0.06%), “15” (1407, 0.06%), “up” (1399,
 0.06%), “blonde” (1398, 0.06%), “day” (1396, 0.06%), “teddy bear” (1386, 0.06%),
 “glass” (1379, 0.06%), “20” (1365, 0.05%), “beer” (1345, 0.05%), “car” (1331, 0.05%),
 “sitting” (1328, 0.05%), “boat” (1326, 0.05%), “standing” (1326, 0.05%), “clear”
 (1318, 0.05%), “13” (1318, 0.05%), “nike” (1293, 0.05%), “sand” (1282, 0.05%),
 “open” (1279, 0.05%), “cows” (1271, 0.05%), “bike” (1267, 0.05%), “chocolate” (1266,
 0.05%), “donut” (1263, 0.05%), “airplane” (1247, 0.05%), “birthday” (1241, 0.05%),
 “carrots” (1239, 0.05%), “skis” (1220, 0.05%), “girl” (1220, 0.05%), “many” (1211,
 0.05%), “zoo” (1204, 0.05%), “suitcase” (1199, 0.05%), “old” (1180, 0.05%), “chair”
 (1174, 0.05%), “beige” (1170, 0.05%), “ball” (1169, 0.05%), “ocean” (1168, 0.05%),
 “sandwich” (1168, 0.05%), “tie” (1166, 0.05%), “horses” (1163, 0.05%), “palm”
 (1163, 0.05%), “stripes” (1155, 0.05%), “fall” (1146, 0.05%), “cheese” (1142, 0.05%),
 “scissors” (1134, 0.05%), “round” (1125, 0.05%), “chinese” (1123, 0.05%), “knife”
 (1120, 0.05%), “14” (1110, 0.04%), “toilet” (1099, 0.04%), “don’t know” (1085,
 0.04%), “snowboard” (1083, 0.04%), “truck” (1076, 0.04%), “boy” (1070, 0.04%),
 “coffee” (1070, 0.04%), “cold” (1064, 0.04%), “fruit” (1064, 0.04%), “walking” (1053,

0.04%), “wedding” (1051, 0.04%), “lot” (1050, 0.04%), “sunglasses” (1047, 0.04%), “mountains” (1030, 0.04%), “wall” (1009, 0.04%), “elephants” (1006, 0.04%), “wet-suit” (998, 0.04%), “square” (994, 0.04%), “toothbrush” (989, 0.04%), “sleeping” (986, 0.04%), “fire hydrant” (977, 0.04%), “bicycle” (973, 0.04%), “overcast” (968, 0.04%), “donuts” (961, 0.04%), “plastic” (961, 0.04%), “breakfast” (955, 0.04%), “tv” (953, 0.04%), “paper” (952, 0.04%), “ground” (949, 0.04%), “asian” (938, 0.04%), “plaid” (936, 0.04%), “dirt” (933, 0.04%), “mirror” (928, 0.04%), “usa” (928, 0.04%), “chicken” (925, 0.04%), “plate” (920, 0.04%), “clock” (912, 0.04%), “luggage” (908, 0.04%), “none” (908, 0.04%), “street” (905, 0.04%), “on table” (904, 0.04%), “spoon” (899, 0.04%), “cooking” (898, 0.04%), “daytime” (896, 0.04%), “16” (893, 0.04%), “africa” (890, 0.04%), “stone” (884, 0.04%), “not sure” (873, 0.04%), “window” (868, 0.03%), “sun” (865, 0.03%), “gold” (860, 0.03%), “people” (856, 0.03%), “racket” (847, 0.03%), “zebras” (845, 0.03%), “carrot” (841, 0.03%), “person” (835, 0.03%), “fish” (835, 0.03%), “happy” (824, 0.03%), “circle” (822, 0.03%), “oranges” (817, 0.03%), “backpack” (812, 0.03%), “25” (810, 0.03%), “leaves” (809, 0.03%), “watch” (804, 0.03%), “mountain” (800, 0.03%), “no one” (798, 0.03%), “ski poles” (792, 0.03%), “city” (791, 0.03%), “couch” (790, 0.03%), “afternoon” (782, 0.03%), “jeans” (781, 0.03%), “brown and white” (779, 0.03%), “summer” (774, 0.03%), “giraffes” (772, 0.03%), “computer” (771, 0.03%), “refrigerator” (768, 0.03%), “birds” (762, 0.03%), “child” (761, 0.03%), “park” (759, 0.03%), “flying kite” (756, 0.03%), “restaurant” (747, 0.03%), “evening” (738, 0.03%), “graffiti” (736, 0.03%), “30” (730, 0.03%), “grazing” (727, 0.03%), “flower” (723, 0.03%), “remote” (720, 0.03%), “hay” (719, 0.03%), “50” (716, 0.03%).

A.9 Additional Examples

To provide insight into the dataset, we provide additional dataset examples (random selection) in Fig. 44, Fig. 45, and Fig. 46.

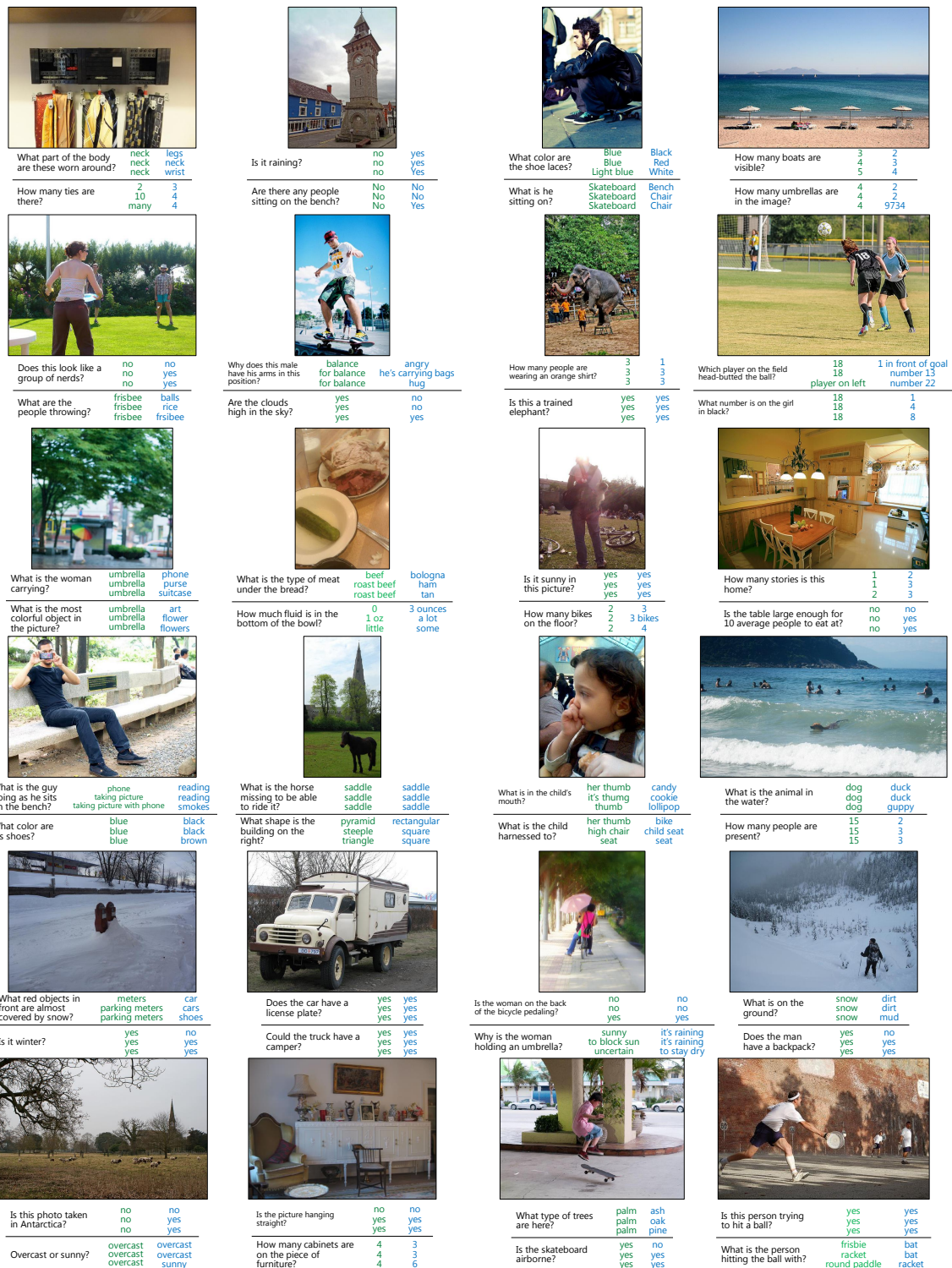


Figure 44: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset.







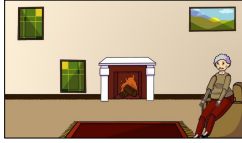
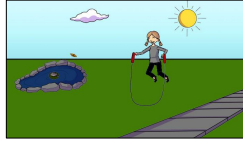
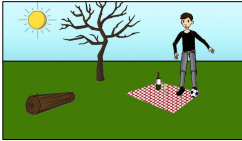

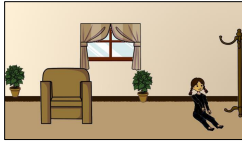

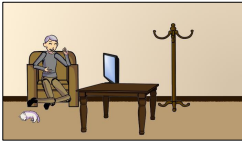




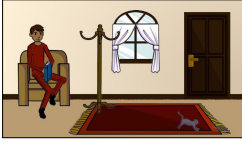
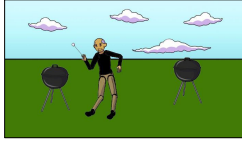
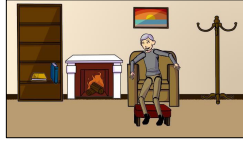
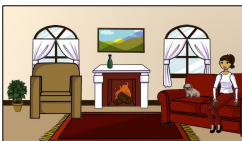
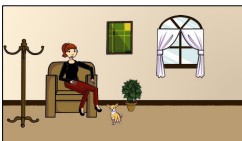

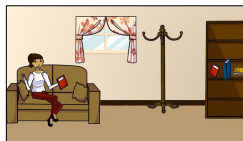
| | | | |
|---|---|---|------------------------|
|  | Who is holding the football? man man boy girl man | cool and sunny mostly sunny partly cloudy | nice sunny sunny |
|  | What is the woman doing? sitting sitting reading reading watching tv | lady woman woman woman women | |
|  | How many bushes are in the background? 3 3 3 4 7 8 | playing playing playing crying eating talking on phone | |
|  | What color is the bike? orange orange orange blue pink red | Is the man injured? no no no no no yes | |
|  | What is the dog looking at? ball soccerball cat cat tree | Will the boy play with the dog? yes yes yes yes yes yes | |
|  | What color is the scooter? red red red red red yellow | How many turtles? 2 2 2 2 3 15 | |
|  | What part of the chair is the lady sitting on? arm arm arm arm seat seat | Is the woman sad? her cat died yes yes no no no | |
|  | What is the little girl playing with? jump rope jump rope jump rope doll dolls teddy bear | What is in the pond? frog lily pad lily pad fish fish turtle | |
|  | Are there leaves in the tree? no no no yes yes yes | What is under the mans left foot? ball soccer ball dollar grass ground | |
|  | What color is the book the woman is reading? blue blue blue blue green red | Is the lady reading? no yes yes no yes yes | |
|  | What is the girl sitting on? floor floor floor floor bench chair rock | What is the girl doing? sitting on floor sitting on floor sit ups dancing singing sleeping | |
|  | What are the boy and girl sitting on? seesaw see saw teeter-totter bench chair couch | What geometric shape is the base of the seesaw? triangle triangle triangle triangle triangle triangle | |
|  | Is the man happy? yes yes yes yes yes yes | Is there an animal in the picture? yes yes yes yes yes yes | |
|  | How many cats are sleeping on the rug? 1 2 4 1 1 2 | What color is the dog on the left? brown brown and white tan and white brown brown brown | |
|  | Is the sun shining? yes yes yes no yes yes | What is in the pond? duck duck duck ducks fish fish | |
|  | Does the man have a good heart? no yes yes yes yes yes | How many rabbits are there? 4 4 4 4 4 4 | |
|  | How many different kinds of fruits are available? 2 2 2 3 4 7 | Which objects needs 2 people in order to work? hands seesaw teeter-totter bandsaw firehose jump rope seesaw | |
|  | Is the cat chasing the mouse? yes yes yes yes yes yes | Is the man sad? no yes yes yes yes yes | |
|  | Is the man young or old? old old oldish old old old | Which grill is the man using? 1 on left left left barbeque gas left 1 | |
|  | Is it a warm night? no no no no yes yes | Is the man happy? my best guess is happy yes yes no yes yes | |
|  | How many windows are in this room? 2 2 2 2 4 8 | Is she waiting on someone? yes yes yes no no yes | |
|  | Is the woman standing? no no no yes yes yes | What is beside the chair? dog dog dog cat table table | |
|  | What color is the plant on the left? green green green green green red | Why is the woman eating a salad rather than pizza? dieting on diet she likes salad dieting overweight she's losing weight | |
|  | How many books are in the shelf? 3 3 3 3 8 23 | What is the person holding? book book notebook phone phone tablet | |

Figure 45: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the abstract scene dataset.



Q: Where is the kid pointing?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) park (l) up (m) floor mat (n) so people don't get wet
 (o) down (p) mom (q) pharos (r) ketchup pickle relish mustard

Q: How many people are in the picture on side of refrigerator?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) 108 mph (l) banana, apple (m) 7 (n) 10 many
 (o) fruit salad (p) full swing (q) 5 (r) vattenfall strom fur gewinner



Q: What sport are they playing?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) tennis (l) bodily functions (m) scissors (n) mississippi and meade
 (o) baseball (p) frisbee (q) soccer (r) its advertising object

Q: What is the man in gray pant's job?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) cop (l) umpire (m) snowflake (n) banker
 (o) chef (p) speedboat (q) 10: 32 (r) males



Q: What is the color of freebee?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) brick (l) peach (m) hill (n) vitamin c
 (o) brown (p) christleton (q) bonsai tree (r) black

Q: How old is the child?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) 6 (l) 12 (m) 10 (n) mechanics
 (o) 5 (p) wait here (q) mad (r) recording studio



Q: Is this person's face painted?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) 4498 (l) not (m) camera film (n) keyboard, mouse, booklet
 (o) stairs (p) n200 (q) public storage (r) pasta, sauce, meat

Q: How many umbrellas are in the photo?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) green
 (k) 20 (l) 54 (m) max payne (n) 62
 (o) 12 (p) dresses (q) 3 to 5 (r) two way traffic



Q: How many of the deer are sleeping?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) 5 (l) left of pond (m) 13 (n) plants and cat
 (o) tree base (p) cement (q) 0 (r) green, blue and yellow

Q: What type of wildlife is this park overrun with?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) eating (l) deer (m) mosquitoes (n) soup
 (o) birds (p) ants (q) girl's (r) woman on right



Q: Where is the blanket?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) fat (l) lying down (m) bed (n) utensils
 (o) on bed (p) grass (q) ground (r) watching child

Q: What is for dessert?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) cake (l) pie (m) a (n) doll and dollhouse
 (o) ice cream (p) yellow book (q) cheesecake (r) there are no fish



Q: Is the girl standing?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) yes! (l) standing (m) hiding (n) sitting
 (o) to sleep (p) bird nest (q) slide (r) park ranger

Q: Does the girl have a lot of toys?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) fork (l) deer (m) rock (n) y
 (o) slide (p) yes 3 of them (q) no image (r) children and toys



Q: Why does the little girl not look happy?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) indian (l) upset (m) dog left (n) smiling at it
 (o) corner (p) to be pet (q) she fell (r) boy is playing with her toys

Q: Why is the boy playing with his sister's toys?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) he likes them (l) parking it (m) dogs (n) shelf
 (o) he feeds them (p) lonely (q) bored (r) likes them



Q: Why are they standing?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) playing game (l) sheepskin (m) waiting (n) no where to sit
 (o) firestone (p) rugby (q) forks (r) waiting for train

Q: Is the TV on?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) shag (l) jeopardy (m) sports (n) between big elephants
 (o) edinburgh (p) strawberries (q) tv show (r) white streak on face



Q: How many legs does the dog have?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) outdoors (l) hiding (m) 45 (n) sitting in grass
 (o) owls (p) 8 (q) 12 (r) arm of sofa

Q: Is the boy at the top of the ladder?

(a) yes (b) no (c) 1 (d) 2 (e) 3 (f) 4
 (g) white (h) red (i) blue (j) yellow
 (k) not sure (l) yellow dog (m) bottom (n) behind trees
 (o) a (p) girl on right (q) top (r) she's in middle

Figure 46: Random examples of multiple-choice questions for numerous representative examples of the real and abstract scene dataset.

APPENDIX B

APPENDIX FOR ANALYZING THE BEHAVIOR OF VQA MODELS

In this appendix, we provide:

1. - Behavioral analysis for question-only and image-only VQA models.
2. - Scatter plot of average distance of test instances from nearest neighbor training instances w.r.t. VQA accuracy.
3. - Additional qualitative examples for “generalization to novel test instances”.
4. - The analyses on “complete question understanding” for different question types.
5. - Additional qualitative examples for “complete question understanding”.
6. - The analyses on “complete image understanding” for different question types.
7. - Additional qualitative examples for “complete image understanding”.

B.1 Behavioral analysis for question-only and image-only VQA models

We evaluated the performance of both CNN+LSTM and ATT models by just feeding in the question (and mean image embedding) and by just feeding in the image (and mean question embedding). We computed the percentage of responses that change on feeding the question as well, compared to only feeding in the image and the percentage of responses that change on feeding the image as well, compared to only feeding in the question. We found that that the responses changed much more (about 40% more) on addition of the question than they did on addition of the image. So this suggests that the VQA models are heavily driven by question rather than the

image.

B.2 Scatter plot of average distance of test instances from nearest neighbor training instances w.r.t. VQA accuracy

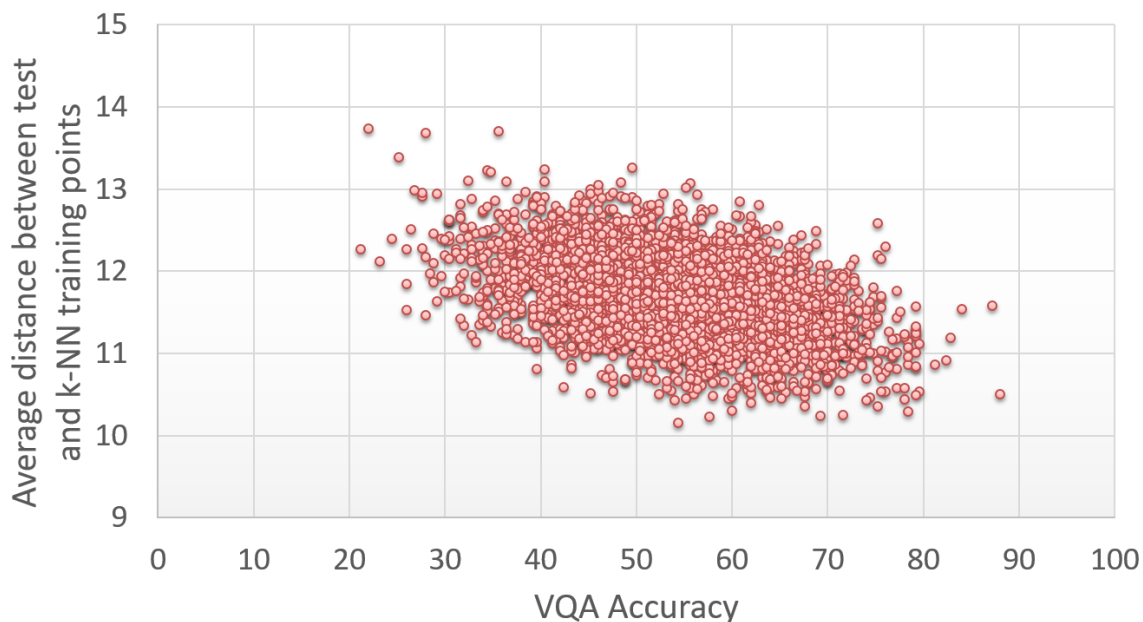


Figure 47: Test accuracy vs. average distance of the test points from k-NN training points for the CNN+LSTM model.

Fig. 47 shows the variation of accuracy of test point w.r.t their average distance from k-NN training points for the CNN+LSTM model. Each point in the plot represents average statistics (accuracy and average distance) for a random subset of 25 test points. We can see that for the test points with low accuracy, the average distance is higher compared to test points with high accuracy. The correlation between accuracy and average distance is significant (-0.41 at $k = 50$.¹)

¹ $k = 50$ leads to highest correlation

B.3 Additional qualitative examples for “generalization to novel test instances”

Fig. 48 shows test QI pairs for which the CNN+LSTM model produces the correct response and their nearest neighbor QI pairs from training set. It can be seen that the nearest neighbor QI pairs from the training set are similar to the test QI pair. In addition, the GT labels in the training set are similar to the test GT label.

Fig. 49 shows test QI pairs for which the CNN+LSTM model produces incorrect response and their nearest neighbor QI pairs from training set. Some of the mistakes are probably because the test QI pair does not have similar QI pairs in the training set (rows 2, 4 and 5) while other mistakes are probably because the GT labels in the training set are not similar to the GT test label (rows 1 and 3).

B.4 Analyses on “complete question understanding” for different question types

We show the breakdown of our analyses from chapter 4 – (i) whether the model ‘listens’ to the entire question; and (ii) which POS tags matter the most – over the three major categories of questions – “yes/no”, “number” and “other” as categorized in [27]. “yes/no” are questions whose answers are either “yes” or “no”, “number” are questions whose answers are numbers (e.g., “Q: How many zebras are there?”, “A: 2”), “other” are rest of the questions.

For “yes/no” questions, the ATT model seems particularly ‘jumpy’ – converging on a predicted answer listening to only the first few words of the question (see Fig. 50). Surprisingly, the accuracy is also as much as the final accuracy (after listening to entire question) when making predictions based on first few words of the question. In contrast, the CNN+LSTM model converges on a predicted answer later, after listening to atleast 35% of the question, achieving as much as the final accuracy after convergence. For “number” and “other” questions, both ATT and CNN+LSTM

model show similar trends (see Fig. 51 for “number” and Fig. 52 for “other”).

It is interesting to note that VQA models are most sensitive to adjectives for “yes/no” questions (compared to wh-words for all questions) (see Fig. 53). This is probably because often the “yes/no” questions are about attributes of objects (e.g., “Is the cup empty?”). For “number” questions, the CNN+LSTM model is most sensitive to adjectives whereas the ATT model is most sensitive to wh-words (see Fig. 54). For “other” questions, both the models are most sensitive to “nouns” (see Fig. 55).

B.5 Additional qualitative examples for “complete question understanding”

Fig. 56 shows examples where the CNN+LSTM model converges on a predicted answer without listening to the entire question. On doing so, the model gets the answer correct for some QI pairs (first three rows) and incorrect for others (last two rows).

B.6 Analyses on “complete image understanding” for different question types

Fig. 57, Fig. 58 and Fig. 59 show the breakdown of percentage of questions for which the model produces same answer across images for “yes/no”, “number” and “other” respectively. The ATT model seems to be more “stubborn” (does not change its answers across images) for “yes/no” questions compared to the CNN+LSTM model, and less “stubborn” for “number” questions compared to the CNN+LSTM model.

B.7 Additional qualitative examples for “complete image understanding”

Fig. 60 shows examples where the CNN+LSTM model produces the same answer for atleast half the images for a given question and the accuracy achieved by the model for such QI pairs.































| Test Sample | Nearest Neighbor Training Samples | | | | |
|--|---|---|--|--|--|
| <p>Q: Does someone have a birthday?</p>  <p>Predicted A: yes GT A: yes Accuracy: 100.0</p> | <p>Q: Could it be someone's birthday?</p>  <p>GT A: yes</p> | <p>Q: Might today be her birthday?</p>  <p>GT A: yes</p> | <p>Q: Does someone have a birthday?</p>  <p>GT A: yes</p> | <p>Q: Is there a basket on the bicycle?</p>  <p>GT A: yes</p> | <p>Q: Is there a balloon on the table?</p>  <p>GT A: yes</p> |
| <p>Q: Which vehicle is towing a small trailer?</p>  <p>Predicted A: motorcycle GT A: motorcycle Accuracy: 100.0</p> | <p>Q: Which vehicle has a picture of a wooly mammoth?</p>  <p>GT A: motorcycle</p> | <p>Q: What is the police officer riding in the picture?</p>  <p>GT A: motorcycle</p> | <p>Q: What type of transportation?</p>  <p>GT A: motorcycle</p> | <p>Q: What is parked next to the motorbike?</p>  <p>GT A: bicycle</p> | <p>Q: What type of transportation is this?</p>  <p>GT A: motorcycle</p> |
| <p>Q: What is the woman doing?</p>  <p>Predicted A: playing wii GT A: playing wii Accuracy: 100.0</p> | <p>Q: What is the woman doing?</p>  <p>GT A: talking on phone</p> | <p>Q: What is the woman doing?</p>  <p>GT A: playing wii</p> | <p>Q: What is the girl doing?</p>  <p>GT A: playing wii</p> | <p>Q: What is this lady doing?</p>  <p>GT A: waiting</p> | <p>Q: What is the woman doing?</p>  <p>GT A: playing wii</p> |
| <p>Q: What color is the sky?</p>  <p>Predicted A: blue GT A: blue Accuracy: 100.0</p> | <p>Q: What color is the sky?</p>  <p>GT A: blue</p> | <p>Q: What color is the sky?</p>  <p>GT A: blue</p> | <p>Q: What color is the sky?</p>  <p>GT A: blue</p> | <p>Q: What color is the sky?</p>  <p>GT A: blue</p> | <p>Q: What color is the sky?</p>  <p>GT A: orange</p> |
| <p>Q: How many tusks does the elephant have?</p>  <p>Predicted A: 2 GT A: 2 Accuracy: 100.0</p> | <p>Q: How many tusks does this animal have?</p>  <p>GT A: 2</p> | <p>Q: How many tusks does the elephant have?</p>  <p>GT A: 2</p> | <p>Q: How many tusks does this elephant have?</p>  <p>GT A: 1</p> | <p>Q: How many tusks does the animal have?</p>  <p>GT A: 2</p> | <p>Q: How many tusks does the elephant has?</p>  <p>GT A: 1</p> |

Figure 48: Test QI pairs for which the CNN+LSTM model produces the correct response and their nearest neighbor QI pairs from training set.































| Test Sample | Nearest Neighbor Training Samples | | | | |
|--|---|--|---|--|---|
| <p>Q: What kind of food is this?</p>  <p>Predicted A: dessert GT A: cereal with fruit Accuracy: 0.0</p> | <p>Q: What kind of food is this?</p>  <p>GT A: dessert</p> | <p>Q: What kind of food is this?</p>  <p>GT A: pizza</p> | <p>Q: What kind of food is this?</p>  <p>GT A: lunch</p> | <p>Q: What type of food is this?</p>  <p>GT A: pizza</p> | <p>Q: What kind of food is this?</p>  <p>GT A: salad</p> |
| <p>Q: What is red and driving down the road?</p>  <p>Predicted A: car GT A: bus Accuracy: 0.0</p> | <p>Q: What is the back of the motorbike?</p>  <p>GT A: box</p> | <p>Q: What is on the pole behind the bike?</p>  <p>GT A: sign</p> | <p>Q: What is around the corner to the right?</p>  <p>GT A: store</p> | <p>Q: What is the bike locked up to?</p>  <p>GT A: tree</p> | <p>Q: What is green and behind the people?</p>  <p>GT A: trees</p> |
| <p>Q: What breed of horse is this?</p>  <p>Predicted A: black and white GT A: clydesdale Accuracy: 0.0</p> | <p>Q: What breed of horse is this?</p>  <p>GT A: brown</p> | <p>Q: What kind of horse is this?</p>  <p>GT A: brown</p> | <p>Q: What kind of horse is this?</p>  <p>GT A: brown</p> | <p>Q: What type of horse is this?</p>  <p>GT A: brown</p> | <p>Q: Which kind of horse is this?</p>  <p>GT A: brown</p> |
| <p>Q: Is this Miley Cyrus?</p>  <p>Predicted A: yes GT A: no Accuracy: 0.0</p> | <p>Q: Is the train blue?</p>  <p>GT A: yes</p> | <p>Q: Does this look right?</p>  <p>GT A: no</p> | <p>Q: Is the skateboard flying?</p>  <p>GT A: yes</p> | <p>Q: Is the bus driver visible?</p>  <p>GT A: no</p> | <p>Q: Is the person female?</p>  <p>GT A: yes</p> |
| <p>Q: What is the name a state that grows these fruits?</p>  <p>Predicted A: new york GT A: florida Accuracy: 0.0</p> | <p>Q: What state is the can from?</p>  <p>GT A: new york</p> | <p>Q: What state is the mug from?</p>  <p>GT A: new york</p> | <p>Q: What face does the topmost fruit have?</p>  <p>GT A: happy</p> | <p>Q: What state is the bear representing?</p>  <p>GT A: new york</p> | <p>Q: What state is the truck from?</p>  <p>GT A: california</p> |

Figure 49: Test QI pairs for which the CNN+LSTM model produces incorrect response and their nearest neighbor QI pairs from training set.

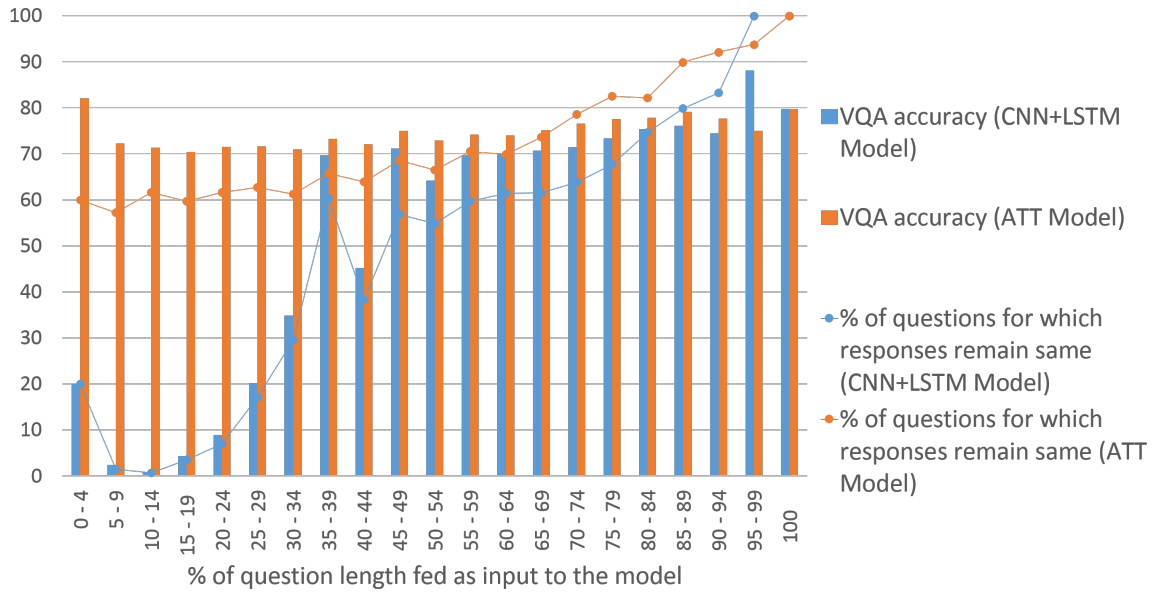


Figure 50: X-axis shows length of partial “yes/no” question (in %) fed as input. Y-axis shows percentage of “yes/no” questions for which responses of these partial “yes/no” questions are the same as full “yes/no” questions and VQA accuracy of partial “yes/no” questions.

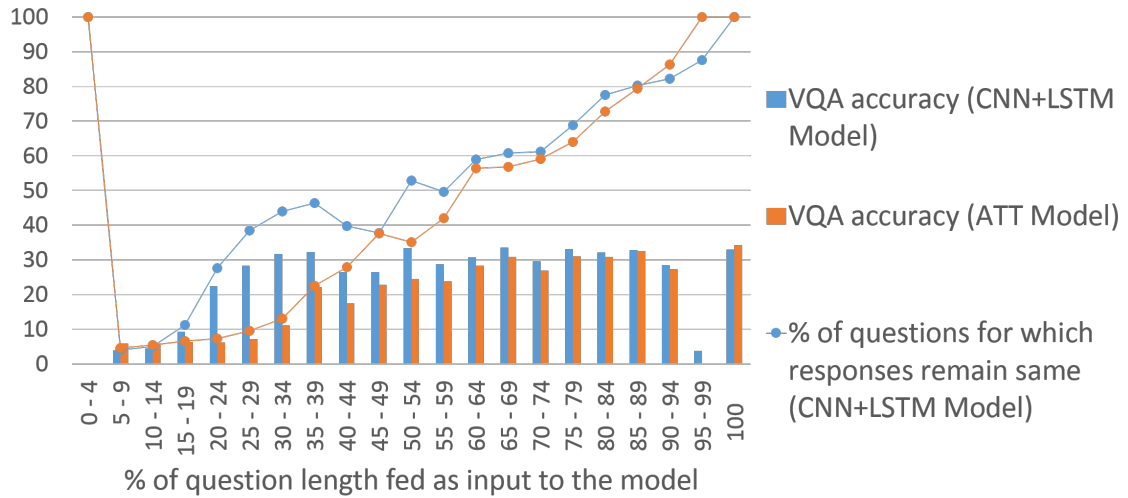


Figure 51: X-axis shows length of partial “number” question (in %) fed as input. Y-axis shows percentage of “number” questions for which responses of these partial “number” questions are the same as full “number” questions and VQA accuracy of partial “number” questions.

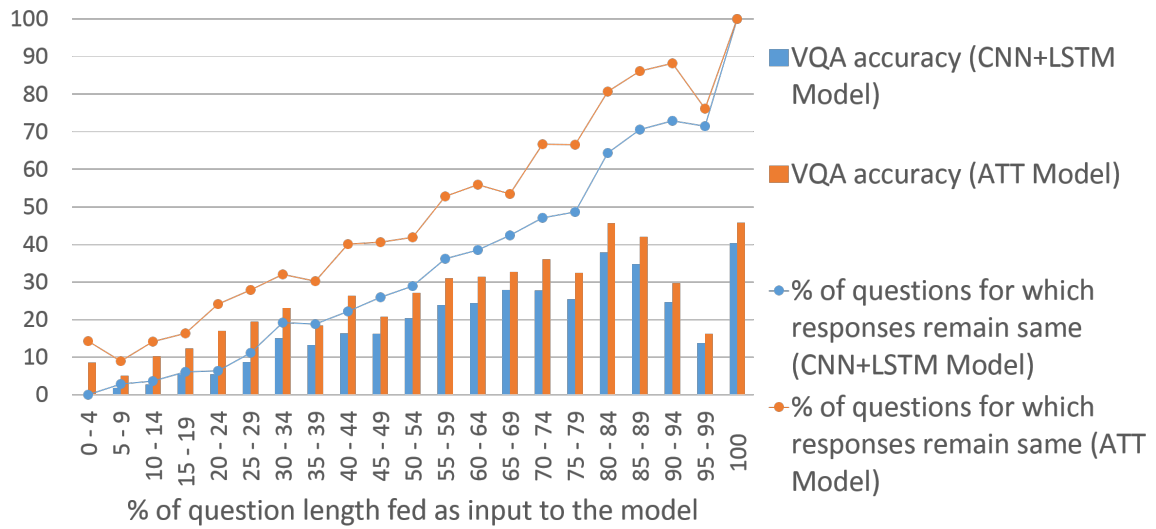


Figure 52: X-axis shows length of partial “other” question (in %) fed as input. Y-axis shows percentage of “other” questions for which responses of these partial “other” questions are the same as full “other” questions and VQA accuracy of partial “other” questions.

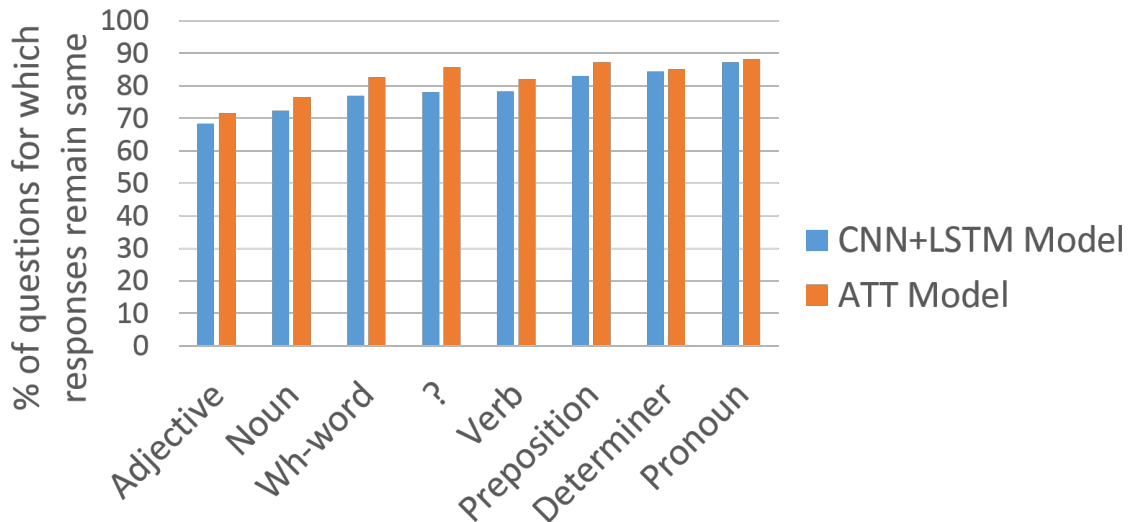


Figure 53: Percentage of “yes/no” questions for which responses remain same (compared to entire “yes/no” question) as a function of POS tags dropped from the “yes/no” question.

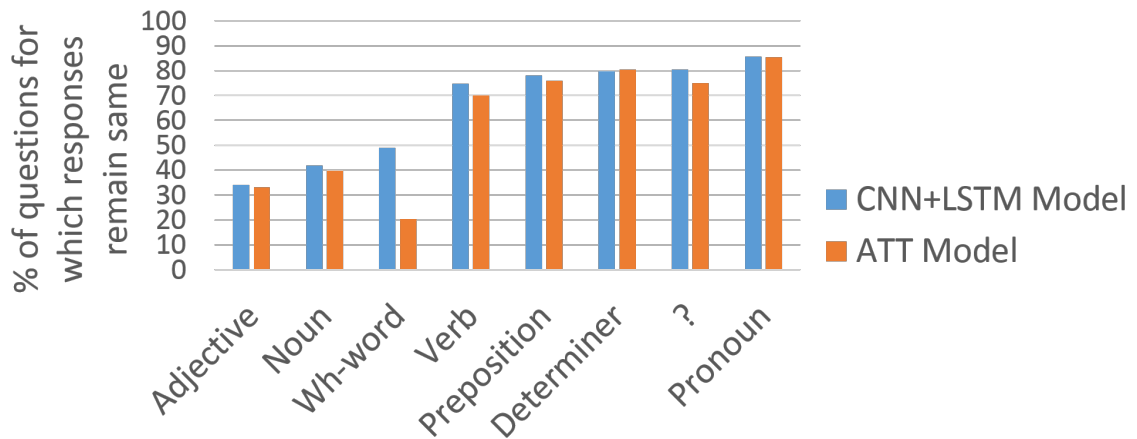


Figure 54: Percentage of “number” questions for which responses remain same (compared to entire “number” question) as a function of POS tags dropped from the “number” question.

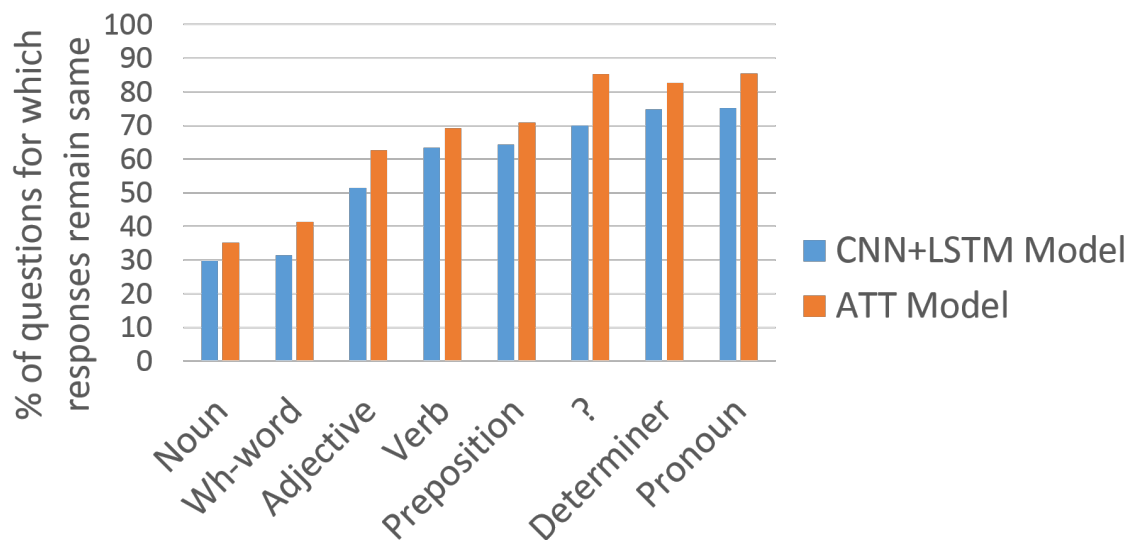


Figure 55: Percentage of “other” questions for which responses remain same (compared to entire “other” question) as a function of POS tags dropped from the “other” question.

| | | |
|---|--|--|
|  | <p>GT A: no</p> <p>Accuracy of predicted answer for full question: 100.0</p> | <p>Q: Is there a tram to the west of where the people are? A: yes</p> <p>Q: Is A: outside</p> <p>Q: Is there A: beach</p> <p>Q: Is there A: yes</p> <p>Q: Is there a tram A: yes</p> <p>Q: Is there a tram to A: yes</p> <p>Q: Is there a tram to the A: yes</p> <p>Q: Is there a tram to the west A: yes</p> <p>Q: Is there a tram to the west of A: yes</p> <p>Q: Is there a tram to the west of where A: yes</p> <p>Q: Is there a tram to the west of where the A: yes</p> <p>Q: Is there a tram to the west of where the people A: yes</p> <p>Q: Is there a tram to the west of where the people are? A: yes</p> <p>Q: Is there a tram to the west of where the people are? A: yes</p> |
|  | <p>GT A: 3</p> <p>Accuracy of predicted answer for full question: 90.0</p> | <p>Q: How many different directions are the benches facing? A: 2</p> <p>Q: How A: yes</p> <p>Q: How many A: 2</p> <p>Q: How many different A: 2</p> <p>Q: How many different directions A: 2</p> <p>Q: How many different directions are A: 2</p> <p>Q: How many different directions are the A: 2</p> <p>Q: How many different directions are the benches A: 2</p> <p>Q: How many different directions are the benches facing? A: 2</p> <p>Q: How many different directions are the benches facing? A: 2</p> |
|  | <p>GT A: grass</p> <p>Accuracy of predicted answer for full question: 100.0</p> | <p>Q: What type of surface is the man standing on? A: grass</p> <p>Q: What A: umbrellas</p> <p>Q: What type A: shadow</p> <p>Q: What type of A: kite</p> <p>Q: What type of surface A: grass</p> <p>Q: What type of surface is A: grass</p> <p>Q: What type of surface is the A: grass</p> <p>Q: What type of surface is the man A: grass</p> <p>Q: What type of surface is the man standing A: grass</p> <p>Q: What type of surface is the man standing on? A: grass</p> <p>Q: What type of surface is the man standing on? A: grass</p> |
|  | <p>GT A: bathroom</p> <p>Accuracy of predicted answer for full question: 0.0</p> | <p>Q: Where is the light fixture in the photo? A: window</p> <p>Q: Where A: bathroom</p> <p>Q: Where is A: outside</p> <p>Q: Where is the A: bathroom</p> <p>Q: Where is the light A: counter</p> <p>Q: Where is the light fixture A: on left</p> <p>Q: Where is the light fixture in A: window</p> <p>Q: Where is the light fixture in the A: window</p> <p>Q: Where is the light fixture in the photo? A: window</p> <p>Q: Where is the light fixture in the photo? A: window</p> |
|  | <p>GT A: continental airlines</p> <p>Accuracy of predicted answer for full question: 0.0</p> | <p>Q: What company is a sponsor of this match? A: polo</p> <p>Q: What A: shadow</p> <p>Q: What company A: nike</p> <p>Q: What company is A: polo</p> <p>Q: What company is a A: polo</p> <p>Q: What company is a sponsor A: polo</p> <p>Q: What company is a sponsor of A: polo</p> <p>Q: What company is a sponsor of this A: polo</p> <p>Q: What company is a sponsor of this match? A: polo</p> <p>Q: What company is a sponsor of this match? A: polo</p> |

Figure 56: Examples where the CNN+LSTM model converges on a predicted answer without listening to the entire question.

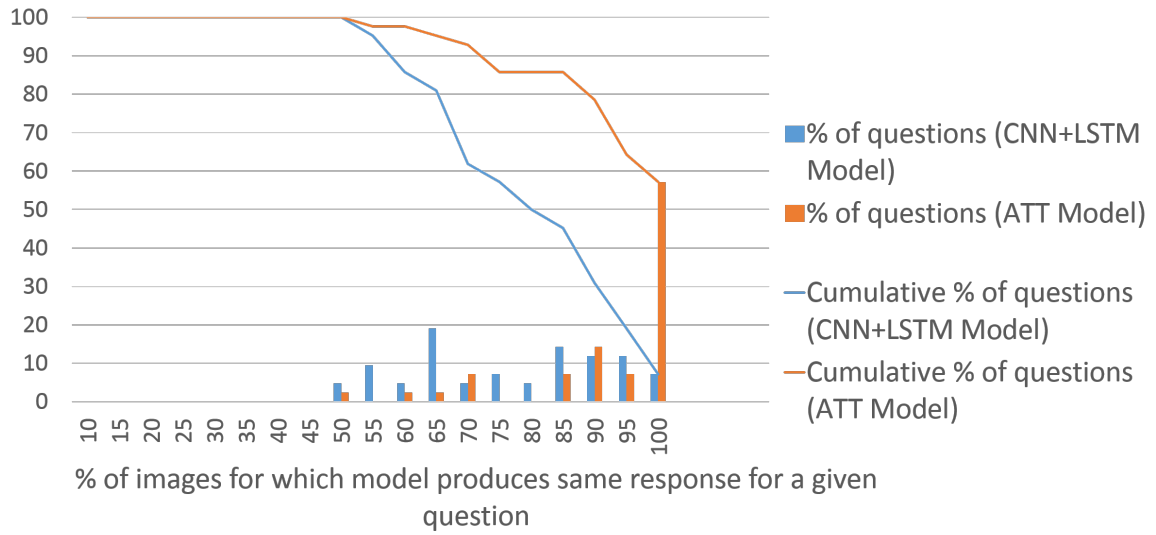


Figure 57: Histogram of percentage of images for which model produces same answer for a given “yes/no” question. The cumulative plot shows the % of “yes/no” questions for which model produces same answer for *atleast* x % of images.

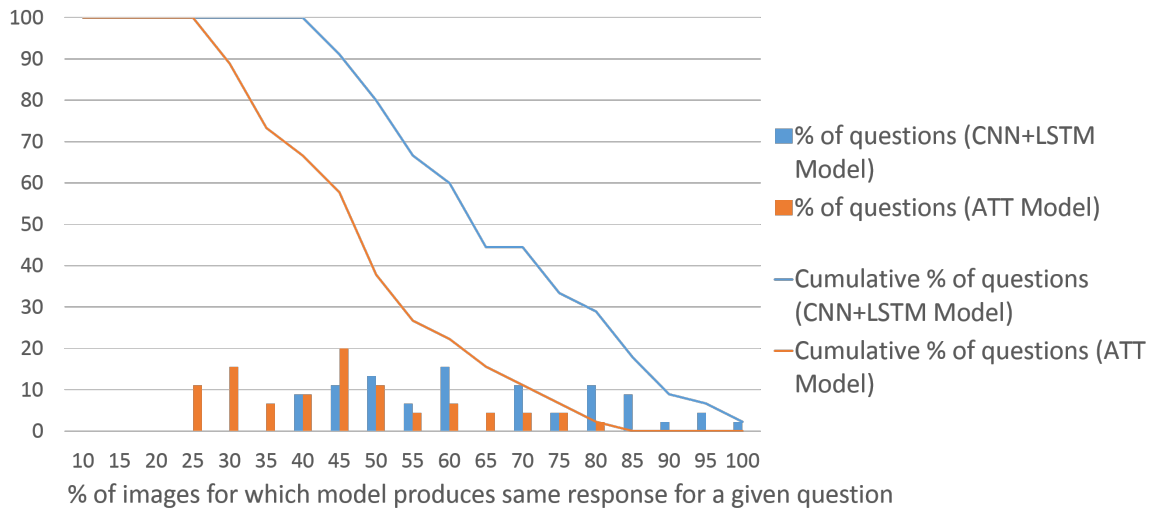


Figure 58: Histogram of percentage of images for which model produces same answer for a given “number” question. The cumulative plot shows the % of “number” questions for which model produces same answer for *atleast* x % of images.

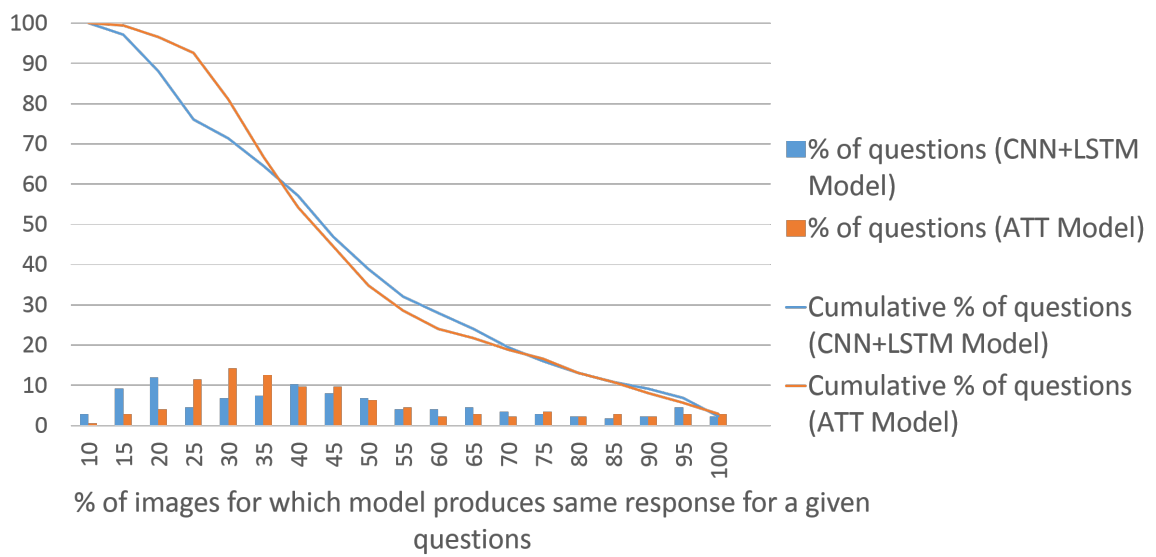


Figure 59: Histogram of percentage of images for which model produces same answer for a given “other” question. The cumulative plot shows the % of “other” questions for which model produces same answer for *atleast* x % of images.

























| | | | | | |
|---|---|---|---|---|---|
| <p>Q: What time is on the clock?</p> <p>A: noon</p> <p>Average Accuracy: 0.0</p> <p>Number of Images: 56</p> |  |  |  |  |  |
| <p>Q: Where is the bus going?</p> <p>A: nowhere</p> <p>Average Accuracy: 3.87</p> <p>Number of Images: 31</p> |  |  |  |  |  |
| <p>Q: What color is the court?</p> <p>A: blue</p> <p>Average Accuracy: 60.29</p> <p>Number of Images: 68</p> |  |  |  |  |  |
| <p>Q: Is the window open?</p> <p>A: yes</p> <p>Average Accuracy: 64.0</p> <p>Number of Images: 35</p> |  |  |  |  |  |
| <p>Q: Is it day or night?</p> <p>A: day</p> <p>Average Accuracy: 96.15</p> <p>Number of Images: 26</p> |  |  |  |  | |

Figure 60: Examples where the CNN+LSTM model produces the same answer for atleast half the images for each of the questions shown above. “Q” denotes the question for which model produces same response for atleast half the images, “A” denotes the answer predicted by the model (which is same for atleast half the images), “Number of Images” denotes the number of images for which the question is repeated in the VQA validation set and “Average Accuracy” is the VQA accuracy for these QI pairs (with same question but different images).

APPENDIX C

APPENDIX FOR OVERCOMING PRIORS IN VQA

C.1 Visual Question Answering under Changing Priors (VQA-CP)

In this appendix, we provide:

1. - Additional analysis of VQA-CP splits
2. - Details of benchmarking VQA models on VQA-CP

C.1.1 Additional analysis of VQA-CP splits

Fig. 61 shows the distribution of answers for several question types such as ‘*what color*’, ‘*what sport*’, ‘*how many*’, etc. for the train (left) and test (right) splits of the VQA-CP v2 dataset (the distribution of answers for VQA-CP v1 is presented in Section 5.1.2). We can see that the distributions of answers for a given question type is significantly different for train and test. For instance, ‘*tennis*’ is the most frequent answer for the question type ‘*what sport*’ in VQA-CP v2 train split whereas ‘*baseball*’ is the most frequent answer for the same question type in VQA-CP v2 test split. Similar differences can be seen for other question types as well – ‘*does*’, ‘*which*’.

C.1.2 Details of benchmarking VQA models on VQA-CP

Below we provide brief descriptions of all the existing VQA models used for benchmarking on VQA-CP splits:

Deeper LSTM Question (d-LSTM Q) [27]: Predicting the answer using question alone (“blind” model). It encodes the question using an LSTM and passes the encoding through a Multi-Layered Perceptron (MLP) to classify into answers.

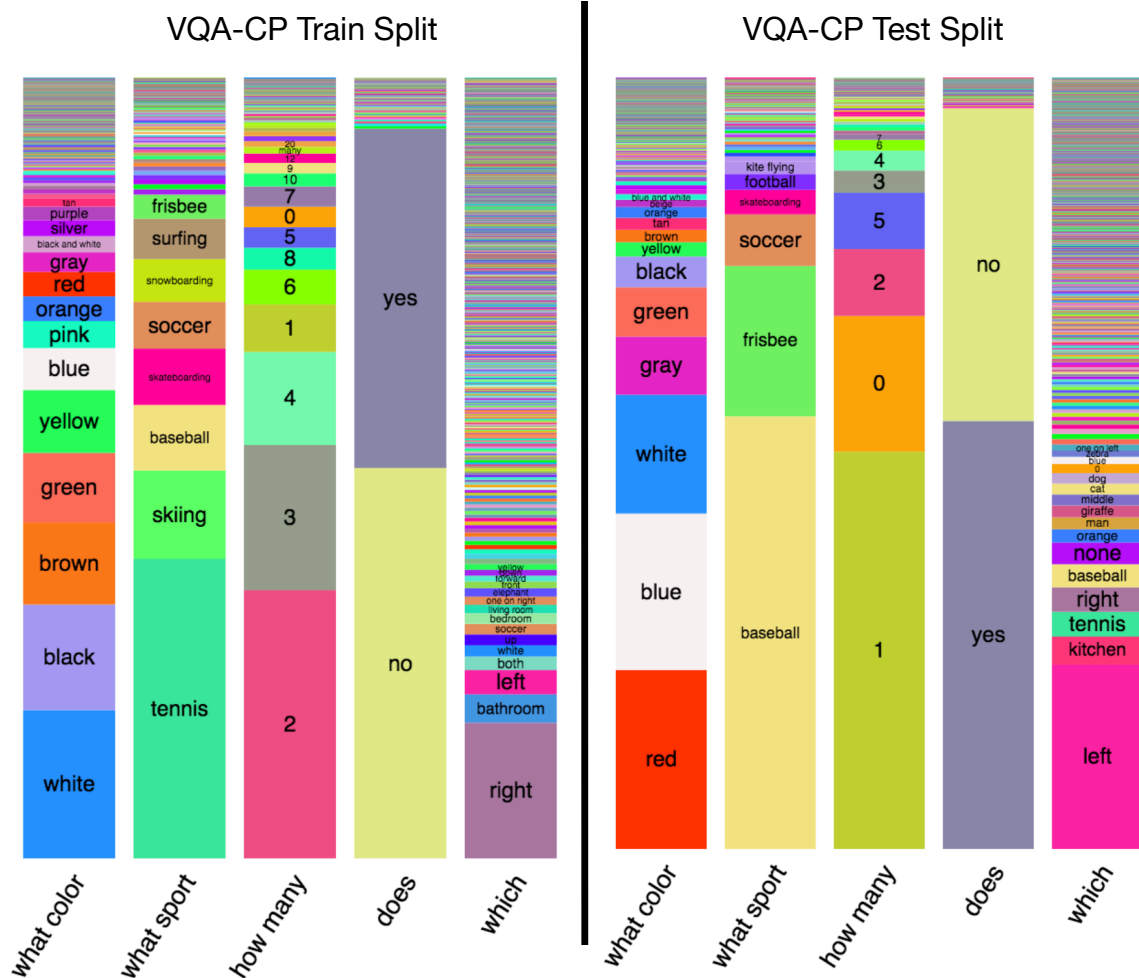


Figure 61: Distribution of answers per question type vary significantly between VQA-CP v2 train (left) and test (right) splits. For instance, ‘white’ and ‘black’ are commonly seen answers in train for ‘What color’, where as ‘red’ is the most frequent answer in test. These have been computed for a random sample of 60K questions.

Deeper LSTM Question + normalized Image (d-LSTM Q + norm I) [27]:

The baseline VQA model. This model consists of a Multi-Layered Perceptron (MLP) fed in by normalized image embeddings (produced by VGG-Net [410]) and question embeddings (produced by a 2 layered LSTM). The MLP produces a distribution over top 1000 answers.

Neural Module Networks (NMN) [24]: The model designed to be compositional in nature. The model consists of composable modules where each module has a specific role (such as detecting a dog in the image, counting the number of dogs

in the image, etc.). Given an image and the natural language question about the image, NMN decomposes the question into its linguistic substructures using a parser to determine the structure of the network required to answer the question.

Stacked Attention Networks (SAN) [495]: One of the widely used models for VQA. Given an image and question, SAN uses the question to attend over the image, using a multi-hop architecture.¹

Multimodal Compact Bilinear Pooling (MCB) [150]: The winner of the VQA Challenge (on real images) 2016. MCB uses multimodal compact bilinear pooling to predict attention over image features and also to combine the attended image features with the question features.

Question-type trends of model performance on VQA-CP : Examining the accuracies of the above VQA models for different question types shows that the performance drop from VQA to VQA-CP is larger for some question types than the others. For VQA-CP v1, all the models show a significant drop ($\sim 70\%$) for ‘*is there a*’ questions (such as ‘*Is there a flowering tree in the scene?*’). For such questions in the VQA-CP v1 test split, the correct answer is ‘*yes*’ whereas the prior answer for questions starting with ‘*Is there a*’ in VQA-CP v1 train split is ‘*no*’. So, models tend to answer the VQA-CP v1 test questions with ‘*no*’ driven by the prior in the training data. Some other examples of question types in VQA-CP v1 resulting in significant drop in performance (more than 10%) for all the models are – ‘*is this an*’, ‘*do you*’, ‘*are there*’, ‘*how many people are*’, ‘*what color is the*’, ‘*what sport is*’, ‘*what room is*’, etc. Examples of question types in VQA-CP v2 resulting in more than 10% drop in performance for all the models are – ‘*is it*’, ‘*is he*’, ‘*are there*’, ‘*how many people are in*’, ‘*what color is the*’, ‘*what animal is*’, ‘*what is in the*’, etc.

¹We use a torch implementation of SAN, available at <https://github.com/abhshkdz/neural-vqa-attention>, for our experiments.

C.2 Grounded Visual Question Answering (GVQA)

In this appendix, we provide:

1. - Implementation details of GVQA
2. - Additional splits of VQA-CP v2
3. - Performance of model components on VQA-CP v2
4. - Performance of SAN with Q_{main}
5. - Performance of GVQA - VCC_{loss} on VQA v1 and VQA v2
6. - Additional qualitative examples

C.2.1 Implementation details of GVQA

For the Question Classifier, we use a single layer LSTM with $512d$ hidden state and train it using the binary cross-entropy loss. For the Answer Cluster Predictor (ACP), we use a single layer LSTM with $256d$ hidden state and train it using the cross-entropy loss (cross-entropy over 50 classes, corresponding to 50 answer clusters). For the Visual Concept Classifier (VCC), we use a single layer LSTM with $512d$ hidden state to encode Q_{main} , the VGG-Net [410] to extract the activations of the last pooling layer ($514 \times 14 \times 14$) and the 2-hop attention architecture from SAN [495]. We use the binary cross-entropy loss to train each classifier in the VCC. For a given training instance, only a subset of all concept clusters are activated, and only these activated clusters contribute to the loss.

For the Question classifier, the ACP and the VCC, we use the rmsprop optimizer with a base learning rate of $3e-4$. For the Answer Predictor (AP) and the Visual Verifier (VV), we use the Adam optimizer with a base learning rate of $3e-3$ and $3e-4$ respectively. All the implementation is using the torch deep learning framework [97].

Effect of number of clusters, clustering algorithm, POS tagger: As mentioned in Section 5.2.2, we used 50 clusters and K-means clustering algorithm for clustering the answer classes for the Answer Cluster Predictor (ACP). We tried 25

and 100 clusters as well and found that changing the number of clusters in K-means from 50 to 25 results in 1.05% drop, from 50 to 100 results in 0.76% drop in the overall VQA accuracy for the VQA-CP v2 dataset. We also tried Agglomerative clustering (instead of K-means) and found that it results in 0.42% drop in the overall VQA accuracy on the VQA-CP v2 dataset. Finally, we tried using Spacy POS tagger (instead of NLTK) for the Concept Extractor (CE) and found that it results in 0.02% improvement in the overall VQA accuracy on the VQA-CP v2 dataset.

C.2.2 Additional splits of VQA-CP v2

| Model | Split1 | Split2 | Split3 | Split4 | Average |
|-------------|--------|--------|--------|--------|---------|
| SAN | 24.96 | 26.07 | 22.69 | 24.19 | 24.48 |
| GVQA | 31.30 | 32.40 | 33.78 | 28.99 | 31.62 |

Figure 62: Performance of SAN and GVQA for different VQA-CP v2 splits. GVQA consistently outperforms SAN across all splits.

As mentioned in Section 5.2.3, to check if our particular VQA-CP split was causing some irregularities in performance, we created three additional sets of VQA-CP v2 splits with different random seeds. We evaluated both SAN and GVQA on all four splits (please see Fig. 62). We can see that GVQA consistently outperforms SAN across all four splits with average improvement being 7.14% (standard error: 1.36).

C.2.3 Performance of model components on VQA-CP v2

Question Classifier: On the VQA-CP v2 test set, the LSTM based question classifier obtains 99.30% accuracy. *ACP:* The Top-1 test accuracy is 51.33%, with 80.12% for questions whose answers are in attribute clusters and 39.21% for questions whose answers are in object clusters. The Top-3 accuracy rises to 63.22%. Note that these accuracies are computed using the automatically created clusters. *VCC:* The weighted mean test F1 score across all classifiers is 0.53. The individual concepts are weighted as per the number of positive samples, reflecting the coverage of that concept in the test set.

C.2.4 Performance of SAN with Q_{main}

Table 14: Performance of SAN - $Q_{full} + Q_{main}$ compared to SAN and GVQA (our model) on VQA-CP v2 dataset. GVQA outperforms both SAN and SAN - $Q_{full} + Q_{main}$.

| Model | Overall | Yes/No | Number | Other |
|-----------------------------|--------------|--------------|--------------|--------------|
| SAN [495] | 24.96 | 38.35 | 11.14 | 21.74 |
| SAN - $Q_{full} + Q_{main}$ | 26.32 | 44.73 | 09.46 | 21.29 |
| GVQA (Ours) | 31.30 | 57.99 | 13.68 | 22.14 |

As mentioned in Section 5.2.4, as an additional check, we trained a version of SAN where the input is Q_{main} instead of Q_{full} . Table C.2.4 shows the results of this version of SAN (SAN - $Q_{full} + Q_{main}$) along with those of SAN and GVQA on VQA-CP v2. We can see that this version of SAN performs 1.36% (overall) better than the original SAN, however still 4.98% (overall) worse than GVQA (with Q_{main}).

C.2.5 Performance of GVQA - VCC_{loss} on VQA v1 and VQA v2

Table 15: Results of GVQA, GVQA - VCC_{loss} and SAN on VQA v1 val split when trained on the VQA v1 train split. Please see text for more details.

| Model | VQA v1 | | | |
|--|---------|--------|--------|-------|
| | Overall | Yes/No | Number | Other |
| SAN | 55.86 | 78.54 | 33.46 | 44.51 |
| GVQA - VCC_{loss} | 48.51 | 65.59 | 32.67 | 39.71 |
| GVQA | 51.12 | 76.90 | 32.79 | 36.43 |
| Ensemble (SAN, SAN) | 56.56 | 79.03 | 34.05 | 45.39 |
| Ensemble ((GVQA - VCC_{loss}), SAN) | 56.44 | 78.27 | 34.45 | 45.62 |
| Ensemble (GVQA, SAN) | 56.91 | 80.42 | 34.40 | 44.96 |
| Oracle (SAN, SAN) | 60.85 | 83.92 | 39.43 | 48.96 |
| Oracle ((GVQA - VCC_{loss}), SAN) | 64.47 | 90.17 | 42.92 | 50.64 |
| Oracle (GVQA, SAN) | 63.77 | 88.98 | 43.37 | 50.03 |

Table C.2.5 and Table C.2.5 present the full results (i.e., broken down into Yes/No, Number and Other) of three models – GVQA, GVQA- VCC_{loss} and SAN, along with their ensembles and Oracle performances. We can see that GVQA- VCC_{loss} performs

Table 16: Results of GVQA, GVQA - VCC_{loss} and SAN on VQA v2 val split when trained on the VQA v2 train split. Please see text for more details.

| Model | Overall | VQA v2 | | |
|--|---------|--------|--------|-------|
| | | Yes/No | Number | Other |
| SAN | 52.02 | 68.89 | 34.55 | 43.80 |
| GVQA - VCC_{loss} | 48.34 | 66.38 | 31.61 | 39.05 |
| GVQA | 48.24 | 72.03 | 31.17 | 34.65 |
| Ensemble (SAN, SAN) | 52.45 | 69.17 | 34.78 | 44.41 |
| Ensemble ((GVQA - VCC_{loss}), SAN) | 51.79 | 68.59 | 34.44 | 43.61 |
| Ensemble (GVQA, SAN) | 52.96 | 72.72 | 34.19 | 42.90 |
| Oracle (SAN, SAN) | 56.68 | 74.37 | 40.08 | 47.61 |
| Oracle ((GVQA - VCC_{loss}), SAN) | 61.93 | 85.13 | 43.51 | 49.16 |
| Oracle (GVQA, SAN) | 61.96 | 85.65 | 43.76 | 48.75 |

worse than GVQA on VQA v1 and similar to GVQA on VQA v2. So in addition to interpretability, GVQA is overall better than GVQA- VCC_{loss} on these original VQA splits. Another observation about GVQA- VCC_{loss} is that the Oracle ((GVQA- VCC_{loss}), SAN)’s overall performance is 8.61% higher than that of SAN for VQA v1 (9.91% for VQA v2), suggesting that GVQA- VCC_{loss} has strengths complementary to SAN (just like GVQA). Note that Oracle ((GVQA- VCC_{loss}), SAN) is higher than Oracle (SAN, SAN) for both VQA v1 and VQA v2, suggesting that GVQA- VCC_{loss} ’s complementary strengths are more than that of another SAN model (with a different random initialization). Inspired by this, we report the performance of the ensemble of GVQA- VCC_{loss} and SAN ((GVQA- VCC_{loss}) + SAN) in Table C.2.5 and Table C.2.5, where the ensemble combines the outputs from the two models using product of confidences of each model. Unlike GVQA + SAN, (GVQA- VCC_{loss}) + SAN does not outperform SAN + SAN (worse by 0.12% overall for VQA v1 and by 0.66% overall for VQA v2). Hence, GVQA is a better complement of SAN than GVQA- VCC_{loss} , in addition to being more transparent.



Figure 63: VCC’s attention map for the example shown in Fig. 25 (right)

C.2.6 Additional qualitative examples

Fig. 63 shows the VCC’s attention map for the example shown in Fig. 25 (right). Please refer to Fig. 25 for more details.

Fig. 64 and Fig. 65 show some qualitative examples from the VQA-CP v2 test set along with GVQA’s and SAN’s predicted answers. Also shown are the intermediate outputs from GVQA which provide insights into why GVQA is predicting what it is predicting and hence enable a system designer to identify the causes of error. This is not easy to do in existing VQA models such as SAN.

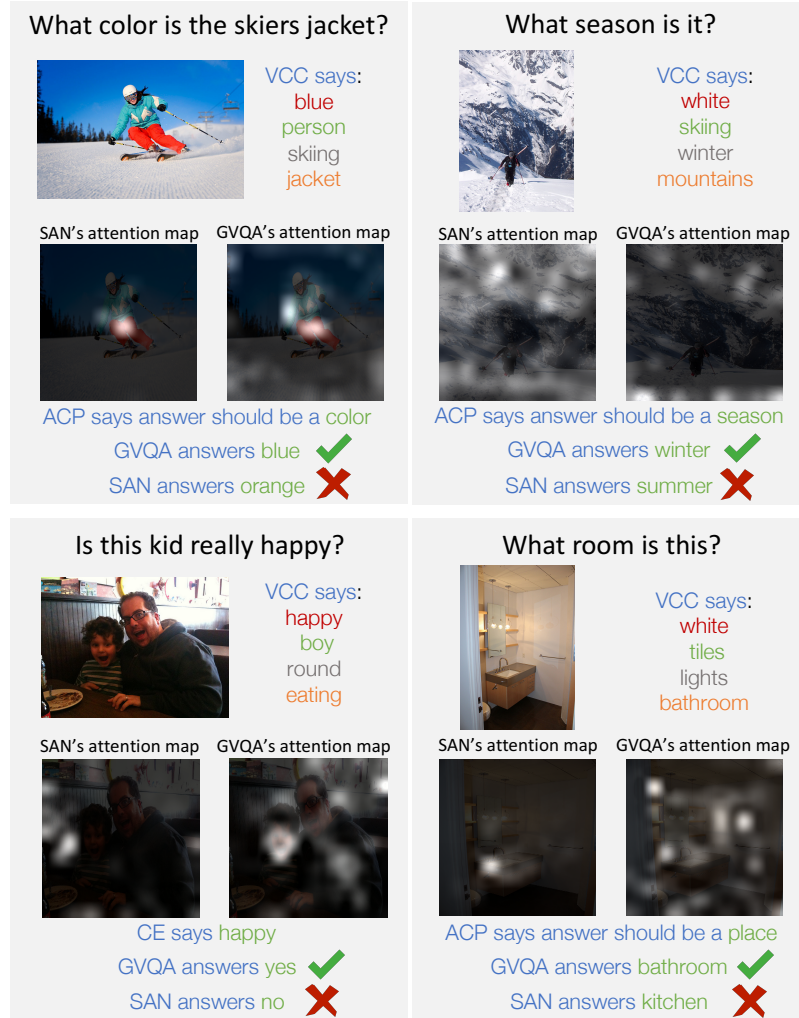


Figure 64: Transparency of GVQA. For each of the above examples, GVQA’s intermediate predictions can help explain why it predicted what it predicted. **Top-left:** VCC predicts the following visual concepts – blue, person, skiing and jacket. ACP predicts the cluster corresponding to colors. Finally, GVQA predicts ‘blue’ as the answer. So, we can see why GVQA predicts ‘blue’ – because, of all the visual concepts predicted by VCC, only ‘blue’ represents a color. Looking at the attention maps can further indicate why GVQA is “seeing” blue (because it is “looking” at the jacket as well, unlike SAN which is only “looking” at the pants). SAN’s prediction is ‘orange’ and unlike GVQA, SAN’s architecture does not facilitate producing such an explanation, which makes it difficult to understand why it is saying what it is saying. **Top-right:** Both GVQA and SAN are “looking” at the regions covered with snow, but GVQA correctly predicts ‘winter’, whereas SAN incorrectly predicts ‘summer’ which is unclear why. **Bottom-left:** The Concept Extractor (CE) predicts ‘happy’ whose visual presence is verified by VCC which is “looking” at the region corresponding to the kid’s face. **Bottom-right:** GVQA focuses on a larger part of the scene and correctly recognizes it as ‘bathroom’.

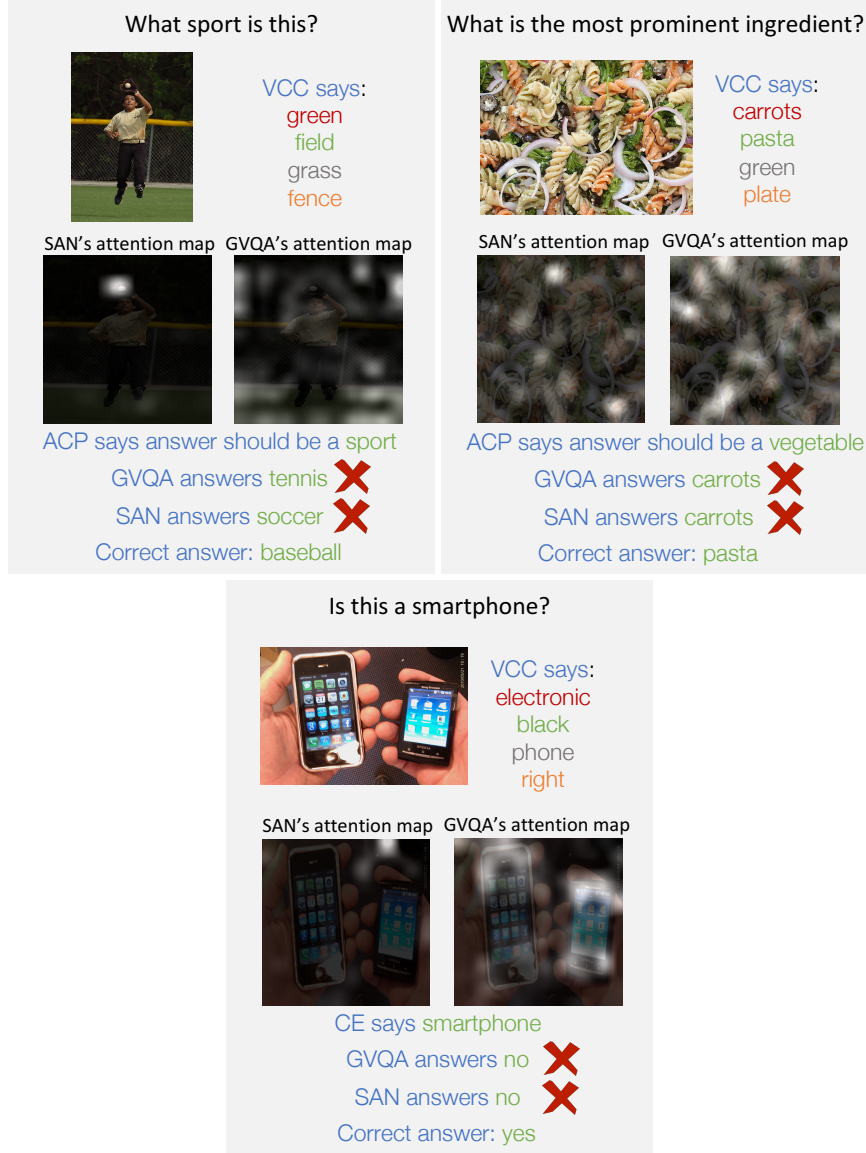


Figure 65: Transparency of GVQA. For the above examples, both GVQA and SAN incorrectly answer the question. However, GVQA’s intermediate predictions can help explain why it is incorrect. **Top-left:** For GVQA, VCC’s predictions indicate that it is perhaps “looking” at the field, which can be further verified by the attention map. SAN’s attention map suggests that it is “looking” at the ball but still does not explain why it is predicting ‘soccer’. Perhaps, it is confusing the ball with a soccer ball. **Top-right:** The attention maps from GVQA and SAN look similar to each other. However, looking at ACP’s and VCC’s prediction (for GVQA) suggest that it is indeed “seeing” ‘pasta’ (the correct answer), but still predicting ‘carrots’ because the ACP is incorrectly predicting the cluster corresponding to vegetables instead of the cluster corresponding to pasta. **Bottom:** GVQA is “looking” at the smartphone (unlike SAN), but yet incorrectly answers ‘no’, because the VCC does not recognize the phone as a smartphone. It however correctly predicts ‘phone’, ‘electronic’, ‘black’ and ‘right’.

REFERENCES

- [1] “Holistic scene understanding via multiple structured hypotheses from perception modules.” Website - https://computing.ece.vt.edu/~aish/holistic_CVPR_workshop_poster.pdf.
- [2] “Personal communication,”
- [3] “Stanford Parser.” <http://nlp.stanford.edu:8080/parser/>.
- [4] “Visual Question Answering Challenge.” Website - <http://www.visualqa.org/challenge.html>.
- [5] “Visual Question Answering Challenge Workshop.” Website - <http://visualqa.org/workshop.html>.
- [6] “Visual Question Answering (VQA) - CloudCV: Large scale distributed computer vision as a cloud service.” Website - <http://cloudcv.org/vqa/>.
- [7] “Education at a glance 2009: OECD indicators.” Organization for Economic Cooperation and Development, 2009.
- [8] AGRAWAL, A., “VQA.” <https://github.com/VT-vision-lab/VQA/>, 2015.
- [9] AGRAWAL, A., BATRA, D., and PARIKH, D., “Analyzing the behavior of visual question answering models,” in *EMNLP*, 2016. 3, 4, 54, 55
- [10] AGRAWAL, A., BATRA, D., and PARIKH, D., “Analyzing the behavior of visual question answering models,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1955–1960, 2016.
- [11] AGRAWAL, A., BATRA, D., and PARIKH, D., “Analyzing the behavior of visual question answering models,” 2016.
- [12] AGRAWAL, A., BATRA, D., PARIKH, D., and KEMBHAVI, A., “Don’t just assume; look and answer: Overcoming priors for visual question answering,” *arXiv preprint arXiv:1712.00377*, 2017. xi, 4, 5
- [13] AGRAWAL, A., BATRA, D., PARIKH, D., and KEMBHAVI, A., “Don’t just assume; look and answer: Overcoming priors for visual question answering,” 2017. 81, 82, 83
- [14] AGRAWAL, A., BATRA, D., PARIKH, D., and KEMBHAVI, A., “Dont just assume; look and answer: Overcoming priors for visual question answering,” 2018.

- [15] AGRAWAL, A., KEMBHAVI, A., BATRA, D., and PARIKH, D., “C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset,” *arXiv preprint arXiv:1704.08243*, 2017.
- [16] AGRAWAL, A., KEMBHAVI, A., BATRA, D., and PARIKH, D., “C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset,” *arXiv preprint arXiv:1704.08243*, 2017.
- [17] AGRAWAL, A., KEMBHAVI, A., BATRA, D., and PARIKH, D., “C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset,” 2017.
- [18] AGRAWAL, A., LU, J., ANTOL, S., MITCHELL, M., ZITNICK, C. L., PARIKH, D., and BATRA, D., “VQA: Visual Question Answering,” 1, 2
- [19] AGRAWAL, A., MALINOWSKI, M., HILL, F., ESLAMI, A., VINYALS, O., and KULKARNI, T., “Generating diverse programs with instruction conditioned reinforced adversarial learning,” 2018. 91
- [20] AGRAWAL, H., MATHIALAGAN, C. S., GOYAL, Y., CHAVALI, N., BANIK, P., MOHAPATRA, A., OSMAN, A., and BATRA, D., “Cloudev: Large-scale distributed computer vision as a cloud service,” in *Mobile Cloud Visual Media Computing*, pp. 265–290, Springer International Publishing, 2015. 35
- [21] ANDERSON, P., HE, X., BUEHLER, C., TENNEY, D., JOHNSON, M., GOULD, S., and ZHANG, L., “Bottom-up and top-down attention for image captioning and visual question answering,” 2017. 80, 81, 82, 84
- [22] ANDERSON, P., HE, X., BUEHLER, C., TENNEY, D., JOHNSON, M., GOULD, S., and ZHANG, L., “Bottom-up and top-down attention for image captioning and visual question answering,” in *CVPR*, 2018.
- [23] ANDREAS, J., ROHRBACH, M., DARRELL, T., and KLEIN, D., “Neural module networks,” 2015.
- [24] ANDREAS, J., ROHRBACH, M., DARRELL, T., and KLEIN, D., “Deep compositional question answering with neural module networks,” in *CVPR*, 2016. 3, 12, 44, 55, 58, 59, 66, 125
- [25] ANDREAS, J., ROHRBACH, M., DARRELL, T., and KLEIN, D., “Learning to compose neural networks for question answering,” in *NAACL*, 2016. 3, 12, 44
- [26] ANTOL, S., AGRAWAL, A., LU, J., MITCHELL, M., BATRA, D., LAWRENCE ZITNICK, C., and PARIKH, D., “Vqa: Visual question answering,” pp. 2425–2433, 2015.
- [27] ANTOL, S., AGRAWAL, A., LU, J., MITCHELL, M., BATRA, D., ZITNICK, C. L., and PARIKH, D., “VQA: Visual Question Answering,” in *International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3, 4, 44, 45, 54, 55, 57, 58, 59, 66, 69, 74, 75, 77, 81, 82, 113, 124, 125

- [28] ANTOL, S., CHEN, X., BATRA, T., PARIKH, D., and ZITNICK, C. L., “Abstract Scenes,” *arXiv preprint arXiv:1504.00325*, 2015.
- [29] ANTOL, S., ZITNICK, C. L., and PARIKH, D., “Zero-Shot Learning via Visual Abstraction,” in *ECCV*, 2014. 16, 18, 102
- [30] ARBELAEZ, P., MAIRE, M., FOWLKES, C., and MALIK, J., “Contour detection and hierarchical image segmentation,” *PAMI*, 2011.
- [31] ATZMON, Y., BERANT, J., KEZAMI, V., GLOBERSON, A., and CHECHIK, G., “Learning to generalize to new compositions in image understanding,” *arXiv preprint arXiv:1608.07639*, 2016. 11
- [32] BACH, F., LANCKRIET, G., and JORDAN, M., “Multiple kernel learning, conic duality, and the SMO algorithm,” 2004.
- [33] BACH, N., HUANG, F., and AL-ONAIZAN, Y., “Goodness: A method for measuring machine translation confidence,” in *ACL*, 2011.
- [34] BACHMANN, T., “Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity?,” *European Journal of Cognitive Psychology*, 1991.
- [35] BAHDANAU, D., HILL, F., LEIKE, J., HUGHES, E., KOHLI, P., and GREFFENSTETTE, E., “Learning to follow language instructions with adversarial reward induction,” *arXiv preprint arXiv:1806.01946*, 2018.
- [36] BAKHSHANDEH, O., BUI, T., LIN, Z., and CHANG, W., “Proposing plausible answers for open-ended visual question answering,” *CoRR*, vol. abs/1610.06620, 2016.
- [37] BANDURA, A., BARBARANELLI, C., and CAPRARA, G. V., “Self-efficacy beliefs as shapers of childrens aspirations and career trajectories,” *Child Development*, vol. 72, no. 1, pp. 187–206, 2001.
- [38] BANERJEE, S. and LAVIE, A., “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65–72, 2005.
- [39] BANSAL, A., FARHADI, A., and PARIKH, D., “Towards Transparent Systems: Semantic Characterization of Failure Modes,” in *European Conference on Computer Vision (ECCV)*, 2014.
- [40] BARROW, H. G. and TENENBAUM, J. M., “Recovering intrinsic scene characteristics from images,” in *Computer Vision Systems*, 1978.
- [41] BARROW, H. G. and TENENBAUM, J. M., “Interpreting line drawings as three-dimensional surfaces,” *AI*, 1981.

- [42] BARROW, H. G. and TENENBAUM, J., “Interpreting line drawings as three-dimensional surfaces,” *Artificial Intelligence*, vol. 17, no. 75–116, 1981.
- [43] BATRA, D., GALLAGHER, A. C., PARIKH, D., and CHEN, T., “Beyond trees: Mrf inference via outer-planar decomposition,” in *CVPR*, 2010.
- [44] BATRA, D., SUKTHANKAR, R., and CHEN, T., “Learning class-specific affinities for image labelling,” in *CVPR*, 2008.
- [45] BATRA, D., SUKTHANKAR, R., and CHEN, T., “Semi-supervised clustering via learnt codeword distances,” 2008.
- [46] BATRA, D., YADOLLAHPOUR, P., GUZMAN-RIVERA, A., and SHAKHNAROVICH, G., “Diverse M-Best Solutions in Markov Random Fields,” in *ECCV*, 2012.
- [47] BELONGIE, S., MALIK, J., and PUZICHA, J., “Shape matching and object recognition using shape contexts,” vol. 24, no. 4, pp. 509–522, 2002.
- [48] BERG, A. and MALIK, J., “Geometric blur for template matching,” 2001.
- [49] BERG, T., BERG, A., and SHIH, J., “Automatic attribute discovery and characterization from noisy web data,” in *ECCV*, 2010.
- [50] BIEDERMAN, I., MEZZANOTTE, R., and RABINOWITZ, J., “Scene perception: Detecting and judging objects undergoing relational violations,” *Cognitive psychology*, vol. 14, no. 2, 1982.
- [51] BIGHAM, J. P., JAYANT, C., JI, H., LITTLE, G., MILLER, A., MILLER, R. C., MILLER, R., TATAROWICZ, A., WHITE, B., WHITE, S., and YEH, T., “VizWiz: Nearly Real-time Answers to Visual Questions,” in *User Interface Software and Technology*, 2010. 1, 8, 10, 14, 15, 42
- [52] BISWAS, A. and JACOBS, D. W., “Active image clustering: Seeking constraints from humans to complement algorithms,” pp. 2152–2159, IEEE, 2012.
- [53] BISWAS, A. and PARIKH, D., “Simultaneous active learning of classifiers & attributes via relative feedback,” 2013.
- [54] BLANZ, V. and VETTER, T., “A morphable model for the synthesis of 3D faces,” 1999.
- [55] BLUM, A. and MITCHELL, T., “Combining labeled and unlabeled data with co-training,” in *COLT*, 1998.
- [56] BODEN, M. A., *Mind As Machine: A History of Cognitive Science*. Oxford University Press, 2006.

- [57] BOLLACKER, K., EVANS, C., PARITOSH, P., STURGE, T., and TAYLOR, J., “Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge,” in *International Conference on Management of Data*, 2008. 16
- [58] BORDES, A., CHOPRA, S., and WESTON, J., “Question Answering with Subgraph Embeddings,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [59] BORDES, A., WESTON, J., and USUNIER, N., “Open Question Answering with Weakly Supervised Embedding Models,” in *European Conference on Machine Learning (ECML)*, 2014.
- [60] BORDES, A., CHOPRA, S., and WESTON, J., “Question Answering with Subgraph Embeddings,” *CoRR*, vol. abs/1406.3676, 2014.
- [61] BORDES, A., USUNIER, N., CHOPRA, S., and WESTON, J., “Large-scale Simple Question Answering with Memory Networks,” *CoRR*, vol. abs/1506.02075, 2015.
- [62] BOSHRA, M. and BHANU, B., “Predicting performance of object recognition,” *PAMI*, 2000.
- [63] BOURDEV, L., MAJI, S., and MALIK, J., “Describing people: A poselet-based approach to attribute classification,” 2011.
- [64] BOURDEV, L. and MALIK, J., “Poselets: Body part detectors trained using 3D human pose annotations,” 2009.
- [65] BRADY, M. J. and KERSTEN, D., “Bootstrapped learning of novel objects,” *Journal of Vision*, 2003.
- [66] BRANSON, S., WAH, C., BABENKO, B., SCHROFF, F., WELINDER, P., PERONA, P., and BELONGIE, S., “Visual recognition with humans in the loop,” 2010.
- [67] BROOKS, R., GREINER, R., and BINFORD, T., “Model-based three-dimensional interpretation of two-dimensional images,” 1981.
- [68] BURNS, K., “Mental models of line drawings,” *Perception*, 2001.
- [69] BURNS, K., HENDRICKS, L. A., DARRELL, T., and ROHRBACH, A., “Women also snowboard: Overcoming bias in captioning models,” *arXiv preprint arXiv:1803.09797*, 2018.
- [70] BUTLER, D. J., WULFF, J., STANLEY, G. B., and BLACK, M. J., “A naturalistic open source movie for optical flow evaluation,” in *ECCV*, 2012.
- [71] CANNY, J., “A computational approach to edge detection,” *PAMI*, 1986.

- [72] CARLSON, A., BETTERIDGE, J., KISIEL, B., SETTLES, B., JR., E. R. H., and MITCHELL, T. M., “Toward an Architecture for Never-Ending Language Learning,” in *AAAI*, 2010. 16
- [73] CARREIRA, J. and SMINCHISESCU, C., “Constrained parametric min-cuts for automatic object segmentation, release 1.” <http://sminchisescu.ins.uni-bonn.de/code/cpmc/>.
- [74] CARREIRA, J. and SMINCHISESCU, C., “Constrained parametric min-cuts for automatic object segmentation,” 2010.
- [75] CARTER, L., “Why students with an apparent aptitude for computer science dont choose to major in computer science,” in *Proceedings of SIGCSE Technical Symposium on Computer Science Education*, 2006.
- [76] CAVALLO, R. and JAIN, S., “Efficient crowdsourcing contests,” in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2012.
- [77] CAVALLO, R. and JAIN, S., “Winner-take-all crowdsourcing contests with stochastic production,” in *HCOMP*, 2013.
- [78] CHANG, C.-C. and LIN, C.-J., “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [79] CHAO, W., HU, H., and SHA, F., “Being negative but constructively: Lessons learnt from creating better visual question answering datasets,” in *NAACL*, 2018. 11
- [80] CHAO, W., HU, H., and SHA, F., “Cross-dataset adaptation for visual question answering,” in *CVPR*, 2018. 11
- [81] CHAPELLE, O., “Training a support vector machine in the primal,” *Neural Computation*, 2007.
- [82] CHAPPELLE, O., CHI, M., and A, Z., “A continuation method for semi-supervised svms,” in *ICML*, 2006.
- [83] CHEN, H., GALLAGHER, A., and GIROD, B., “What’s in a name? First names as facial attributes,” 2013.
- [84] CHEN, K., WANG, J., CHEN, L., GAO, H., XU, W., and NEVATIA, R., “ABC-CNN: an attention based convolutional neural network for visual question answering,” *CoRR*, vol. abs/1511.05960, 2015. 3, 44
- [85] CHEN, L.-C., PAPANDREOU, G., KOKKINOS, I., MURPHY, K., and YUILLE, A. L., “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” in *ICLR*, 2015.

- [86] CHEN, X., FANG, H., LIN, T.-Y., VEDANTAM, R., GUPTA, S., DOLLÁR, P., and ZITNICK, C. L., “Microsoft COCO Captions: Data Collection and Evaluation Server,” *arXiv preprint arXiv:1504.00325*, 2015. 19
- [87] CHEN, X., FANG, H., LIN, T., VEDANTAM, R., GUPTA, S., DOLLÁR, P., and ZITNICK, C. L., “Microsoft COCO captions: Data collection and evaluation server,” *CoRR*, vol. abs/1504.00325, 2015. 21
- [88] CHEN, X., SHRIVASTAVA, A., and GUPTA, A., “NEIL: Extracting Visual Knowledge from Web Data,” in *ICCV*, 2013. 16
- [89] CHEN, X. and ZITNICK, C. L., “Mind’s Eye: A Recurrent Visual Representation for Image Caption Generation,” in *CVPR*, 2015. 9, 14
- [90] CHOI, J., RASTEGARI, M., FARHADI, A., and DAVIS, L., “Adding unlabeled samples to categories by learned attributes,” 2013.
- [91] CHOULARTON, S., “Early stage detection of speech recognition errors,” 2009.
- [92] CHOWDHURY, S. N., MALINOWSKI, M., BULLING, A., and FRITZ, M., “Contextual media retrieval using natural language queries,” 2016.
- [93] CHRISTIE, G., LADDHA, A., AGRAWAL, A., ANTOL, S., GOYAL, Y., KOCHERSBERGER, K., and BATRA, D., “Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [94] CHUNG, S., CHRISTOUDIAS, C., DARRELL, T., ZINIEL, S., and KALISH, L., “A novel image based tool to reunite children with their families after disasters,” *Academic Emergency Medicine*, vol. 19, no. 11, pp. 1227–1234, 2012.
- [95] COLE, F., DURAND, F., FREEMAN, W., and ADELSON, E., “Interpreting line drawings of smooth shapes,” *Journal of Vision*, 2011.
- [96] COLE, F., SANIK, K., DECARLO, D., FINKELSTEIN, A., FUNKHOUSER, T., RUSINKIEWICZ, S., and SINGH, M., “How well do line drawings depict shape?,” *ACM Transactions on Graphics*, 2009.
- [97] COLLOBERT, R., KAVUKCUOGLU, K., and FARABET, C., “Torch7: A matlab-like environment for machine learning,” in *BigLearn, NIPS Workshop*, 2011. 127
- [98] COOPER, M. C., “Interpretation of line-drawings of complex objects,” *Image and Vision Computing*, 1993.
- [99] COOPER, S., KHATIB, F., MAKEDON, I., L, H., BARBERO, J., BAKER, D., FOGARTY, J., POPOVIC, Z., and FOLDIT PLAYERS, “Analysis of social gameplay macros in the Foldit cookbook,” in *Proceedings of the International Conference on the Foundations of Digital Games*, 2011.

- [100] COPPERSMITH, G. and KELLY, E., “Dynamic wordclouds and vennclouds for exploratory data analysis,” in *ACL Workshop on Interactive Language Learning and Visualization*, 2014. xiv, 94, 95, 96, 97
- [101] COX, I., MILLER, M., MINKA, T., PAPATHOMAS, T., and YIANILOS, P., “The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments,” *IEEE Transactions on Image Processing*, vol. 9, pp. 20–37, 2000.
- [102] DAI, B., FIDLER, S., URTASUN, R., and LIN, D., “Towards diverse and natural image descriptions via a conditional gan,” 2017. 13
- [103] DALAL, N. and TRIGGS, B., “Histograms of oriented gradients for human detection,” in *CVPR*, 2005.
- [104] DALAL, N. and TRIGGS, B., “Histograms of oriented gradients for human detection,” 2005.
- [105] DAS, A., DATTA, S., GKIOXARI, G., LEE, S., PARIKH, D., and BATRA, D., “Embodied Question Answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [106] DAS, A., KOTTUR, S., GUPTA, K., SINGH, A., YADAV, D., MOURA, J. M., PARIKH, D., and BATRA, D., “Visual Dialog,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 74, 77
- [107] DAS, A., KOTTUR, S., MOURA, J. M., LEE, S., and BATRA, D., “Learning cooperative visual dialog agents with deep reinforcement learning,” *arXiv preprint arXiv:1703.06585*, 2017.
- [108] DE MARNEFFE, M.-C., MACCARTNEY, B., MANNING, C. D., and OTHERS, “Generating typed dependency parses from phrase structure parses,” in *Proceedings of LREC*, vol. 6, pp. 449–454, 2006.
- [109] DE VRIES, H., STRUB, F., CHANDAR, S., PIETQUIN, O., LAROCHELLE, H., and COURVILLE, A., “Guesswhat?! visual object discovery through multi-modal dialogue,” *arXiv preprint arXiv:1611.08481*, 2016.
- [110] DELANY, S. J., CUNNINGHAM, P., and DOYLE, D., “Generating estimates of classification confidence for a case-based spam filter,” in *International Conference on Case-based Reasoning*, 2005.
- [111] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., and FEI-FEI, L., “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR*, 2009.
- [112] DENG, J., BERG, A. C., and FEI-FEI, L., “Hierarchical Semantic Indexing for Large Scale Image Retrieval,” in *CVPR*, 2011. 9

- [113] DESAI, C., RAMANAN, D., and FOWLKES, C., “Discriminative models for multi-class object layout,” 2009.
- [114] DHAR, S., ORDONEZ, V., and BERG, T. L., “High level describable attributes for predicting aesthetics and interestingness,” 2011.
- [115] DONAHUE, J. and GRAUMAN, K., “Annotator rationales for visual recognition,” 2011.
- [116] DONAHUE, J., HENDRICKS, L. A., GUADARRAMA, S., ROHRBACH, M., VENUGOPALAN, S., SAENKO, K., and DARRELL, T., “Long-term Recurrent Convolutional Networks for Visual Recognition and Description,” in *CVPR*, 2015. 9, 14
- [117] DONAHUE, J., JIA, Y., VINYALS, O., HOFFMAN, J., ZHANG, N., TZENG, E., and DARRELL, T., “Decaf: A deep convolutional activation feature for generic visual recognition,” *arXiv preprint arXiv:1310.1531*, 2013.
- [118] DOUZE, M., RAMISA, A., and SCHMID, C., “Combining attributes and fisher vectors for efficient image retrieval,” 2011.
- [119] DREDZE, M. and CRAMMER, K., “Confidence-weighted linear classification,” in *ICML*, 2008.
- [120] DRIVER, J. and FRACKOWIAK, R. S., “Neurobiological measures of human selective attention,” *Neuropsychologia*, vol. 39, no. 12, pp. 1257 – 1262, 2001.
- [121] DRUCK, G., SETTLES, B., and MCCALLUM, A., “Active learning by labeling features,” in *EMNLP*, 2009.
- [122] DUAN, K., PARIKH, D., CRANDALL, D., and GRAUMAN, K., “Discovering localized attributes for fine-grained recognition,” 2012.
- [123] DUIN, R. P. W. and TAX, D. M. J., “Classifier Conditional Posterior Probabilities,” in *Joint IAPR International Workshops on Advances in Pattern Recognition*, 1998.
- [124] ELAZARY, L. and ITTI, L., “Interesting objects are visually salient,” *Journal of Vision*, vol. 8, no. 3:3, pp. 1–15, 2008.
- [125] ELLIOTT, D. and KELLER, F., “Comparing Automatic Evaluation Measures for Image Description,” in *ACL*, 2014. 14
- [126] ESPEHOLT, L., SOYER, H., MUNOS, R., SIMONYAN, K., MNIH, V., WARD, T., DORON, Y., FIROIU, V., HARLEY, T., DUNNING, I., and OTHERS, “Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures,” *arXiv preprint arXiv:1802.01561*, 2018.
- [127] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., and ZISSERMAN, A., “The pascal visual object classes (voc) challenge,” *IJCV*, 2010.

- [128] FADER, A., ZETTLEMOYER, L., and ETZIONI, O., “Paraphrase-Driven Learning for Open Question Answering,” in *ACL*, 2013. 9
- [129] FADER, A., ZETTLEMOYER, L., and ETZIONI, O., “Open Question Answering over Curated and Extracted Knowledge Bases,” in *International Conference on Knowledge Discovery and Data Mining*, 2014. 9
- [130] FANG, H., GUPTA, S., IANDOLA, F. N., SRIVASTAVA, R., DENG, L., DOLLÁR, P., GAO, J., HE, X., MITCHELL, M., PLATT, J. C., ZITNICK, C. L., and ZWEIG, G., “From Captions to Visual Concepts and Back,” *CoRR*, vol. abs/1411.4952, 2014.
- [131] FANG, H., GUPTA, S., IANDOLA, F. N., SRIVASTAVA, R., DENG, L., DOLLÁR, P., GAO, J., HE, X., MITCHELL, M., PLATT, J. C., ZITNICK, C. L., and ZWEIG, G., “From Captions to Visual Concepts and Back,” in *CVPR*, 2015. 9, 14
- [132] FARHADI, A., ENDRES, I., and HOIEM, D., “Attribute-centric recognition for cross-category generalization,” 2010.
- [133] FARHADI, A., ENDRES, I., HOIEM, D., and FORSYTH, D., “Describing objects by their attributes,” 2009.
- [134] FARHADI, A., HEJRATI, M., SADEGHI, A., YOUNG, P., RASHTCHIAN, C., HOCKENMAIER, J., and FORSYTH, D., “Every Picture Tells a Story: Generating Sentences for Images,” in *ECCV*, 2010. 9
- [135] FASEL, B. and LUETTIN, J., “Automatic facial expression analysis: a survey,” *Pattern Recognition*, vol. 36, no. 1, 2003.
- [136] FEI-FEI, L., IYER, A., KOCH, C., and PERONA, P., “What do we perceive in a glance of a real-world scene?,” *Journal of Vision*, 2007.
- [137] FEI-FEI, L., VANRULLEN, R., KOCH, C., and PERONA, P., “Rapid natural scene categorization in the near absence of attention,” *PNAS*, 2002.
- [138] FELZENSZWALB, P., GIRSHICK, R., MCALLESTER, D., and RAMANAN, D., “Object detection with discriminatively trained part based models,” vol. 32, no. 9, pp. 1627–1645, 2010.
- [139] FERECATU, M. and GEMAN, D., “Interactive search for image categories by mental matching,” 2007.
- [140] FERGUS, R., BERNAL, H., WEISS, Y., and TORRALBA, A., “Semantic label sharing for learning with many categories,” 2010.
- [141] FERRARI, V. and ZISSERMAN, A., “Learning visual attributes,” 2007.
- [142] FERRARI, V. and ZISSERMAN, A., “Learning visual attributes,” 2007.

- [143] FISCHLER, M. A. and ELSCHLAGER, R. A., “The representation and matching of pictorial structures,” *IEEE Transactions on Computers*, 1973.
- [144] FLICKNER, M., SAWHNEY, H., NILBACK, W., ASHLEY, J., HUANG, Q., DOM, B., GORKANI, M., HAFNER, J., LEE, D., PETKOVIC, D., STEELE, D., and YANKER, P., “Query by Image and Video Content: The QBIC System,” *IEEE Computer*, vol. 28, pp. 23–32, September 1995.
- [145] FLICKR, “Wayan Vota. Photo used under Creative Commons.”
- [146] FOUHEY, D. F. and ZITNICK, C., “Predicting object dynamics in scenes,” in *CVPR*, 2014.
- [147] FOWLKES, C. C., “Measuring the ecological validity of grouping and figure-ground cues,” *Thesis*, 2005.
- [148] FOWLKES, C. C., MARTIN, D. R., and MALIK, J., “Local figureground cues are valid for natural images,” *Journal of Vision*, vol. 7, no. 8, 2007.
- [149] FROME, A., SINGER, Y., SHA, F., and MALIK, J., “Learning globally-consistent local distance functions for shape-based image retrieval and classification,” 2007.
- [150] FUKUI, A., PARK, D. H., YANG, D., ROHRBACH, A., DARRELL, T., and ROHRBACH, M., “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *EMNLP*, 2016. 3, 44, 46, 55, 58, 59, 61, 66, 126
- [151] FUKUI, A., PARK, D. H., YANG, D., ROHRBACH, A., DARRELL, T., and ROHRBACH, M., “Multimodal compact bilinear pooling for visual question answering and visual grounding,” 2016.
- [152] FUKUI, A., PARK, D. H., YANG, D., ROHRBACH, A., DARRELL, T., and ROHRBACH, M., “Multimodal compact bilinear pooling for visual question answering and visual grounding,” 2016.
- [153] GALLEGUILLOS, C., RABINOVICH, A., and BELONGIE, S., “Object categorization using co-occurrence, location and appearance,” 2008.
- [154] GANIN, Y., KULKARNI, T., BABUSCHKIN, I., ESLAMI, S., and VINYALS, O., “Synthesizing programs for images using reinforced adversarial learning,” 2018. 91
- [155] GAO, H., MAO, J., ZHOU, J., HUANG, Z., WANG, L., and XU, W., “Are you talking to a machine? dataset and methods for multilingual image question,” in *Advances in Neural Information Processing Systems*, pp. 2296–2304, 2015.
- [156] GAO, H., MAO, J., ZHOU, J., HUANG, Z., and YUILLE, A., “Are you talking to a machine? dataset and methods for multilingual image question answering,” in *NIPS*, 2015. 9

- [157] GAULIN, S. J. C. and MCBURNEY, D. H., *Evolutionary Psychology*. Prentice Hall, 2003.
- [158] GEMAN, D., GEMAN, S., HALLONQUIST, N., and YOUNES, L., “A Visual Turing Test for Computer Vision Systems,” in *PNAS*, 2014. 8, 15
- [159] GEMAN, D., GEMAN, S., HALLONQUIST, N., and YOUNES, L., “Visual turing test for computer vision systems,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.
- [160] GIBSON, J. J., “Perception of the visual world,” in *Houghton Mifflin*, 1950.
- [161] GOLLAND, P., “Discriminative direction for kernel classifiers,” 2001.
- [162] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDEFARLEY, D., OZAIR, S., COURVILLE, A., and BENGIO, Y., “Generative adversarial nets,” pp. 2672–2680, 2014.
- [163] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDEFARLEY, D., OZAIR, S., COURVILLE, A., and BENGIO, Y., “Generative adversarial nets,” in *NIPS*, 2014.
- [164] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDEFARLEY, D., OZAIR, S., COURVILLE, A., and BENGIO, Y., “Generative adversarial networks,” 2014. 12, 13
- [165] GORDON, J. and DURME, B. V., “Reporting bias and knowledge extraction,” in *Proceedings of the 3rd Workshop on Knowledge Extraction, at CIKM 2013*, 2013. 94
- [166] GORDON, J. and VAN DURME, B., “Reporting bias and knowledge acquisition,” in *Proceedings of the 2013 workshop on Automated knowledge base construction*, pp. 25–30, ACM, 2013.
- [167] GOYAL, Y., KHOT, T., SUMMERS-STAY, D., BATRA, D., and PARIKH, D., “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” 2016. 74, 75, 77, 81
- [168] GOYAL, Y., KHOT, T., SUMMERS-STAY, D., BATRA, D., and PARIKH, D., “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *CVPR*, 2017. 4, 11, 54, 55, 69
- [169] GRADY, L., JOLLY, M.-P., and SEITZ, A., “Segmentation from a box,” in *ICCV*, 2011.
- [170] GRAY, T., “Brain drain in the tech world?,” *Enterprise News*, July 2005.
- [171] GRUBINGER, M., CLOUGH, P., MÜLLER, H., and DESELAERS, T., “The iaprtc-12 benchmark: A new evaluation resource for visual information systems,” in *Int. Workshop OntoImage*, 2006.

- [172] GUADARRAMA, S., KRISHNAMOORTHY, N., MALKARNENKAR, G., VENUGOPALAN, S., MOONEY, R., DARRELL, T., and SAENKO, K., “YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition,” in *ICCV*, December 2013. 9
- [173] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V., and COURVILLE, A. C., “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 5767–5777, 2017.
- [174] GUPTA, A. and DAVIS, L., “Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers,” 2008.
- [175] GUPTA, A., EFROS, A., and HEBERT, M., “Blocks world revisited: Image understanding using qualitative geometry and mechanics,” 2010.
- [176] GURARI, D., LI, Q., STANGL, A. J., GUO, A., LIN, C., GRAUMAN, K., LUO, J., and BIGHAM, J. P., “Vizwiz grand challenge: Answering visual questions from blind people,” *CVPR*, 2018.
- [177] HANSON, A. and RISEMAN, E., “Visions: A computer system for interpreting scenes,” in *Computer Vision Systems*, Academic Press, 1978.
- [178] HANSON, A. R. and RISEMAN, E. M., “VISIONS: A computer system for interpreting scenes,” in *Computer Vision Systems*, Academic Press, 1978.
- [179] HEDAU, V., HOIEM, D., and FORSYTH, D., “Recovering the spatial layout of cluttered rooms,” 2009.
- [180] HEIDER, F. and SIMMEL, M., “An experimental study of apparent behavior,” *The American Journal of Psychology*, 1944.
- [181] HERMANN, K. M., HILL, F., GREEN, S., WANG, F., FAULKNER, R., SOYER, H., SZEPESVARI, D., CZARNECKI, W. M., JADERBERG, M., TEPLYASHIN, D., WAINWRIGHT, M., APPS, C., HASSABIS, D., and BLUNSOM, P., “Grounded language learning in a simulated 3d world,” *CoRR*, vol. abs/1706.06551, 2017.
- [182] HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B., KLAMBAUER, G., and HOCHREITER, S., “Gans trained by a two time-scale update rule converge to a nash equilibrium,” *CoRR*, vol. abs/1706.08500, 2017.
- [183] HINTON, G. E. and PARSONS, L. A., “Frames of reference and mental imagery,” *J. Long and A. Baddeley, editors, Attention and Performance IX*, 1981.
- [184] HOCHREITER, S. and SCHMIDHUBER, J., “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [185] HOCHREITER, S. and SCHMIDHUBER, J., “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

- [186] HODOSH, M., YOUNG, P., and HOCKENMAIER, J., “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,” *JAIR*, 2013. 14
- [187] HOIEM, D., EFROS, A. A., and HEBERT, M., “Automatic photo pop-up,” in *SIGGRAPH*, 2005.
- [188] HOIEM, D., EFROS, A., and HEBERT, M., “Putting objects in perspective,” 2006.
- [189] HOIEM, D., CHODPATHUMWAN, Y., and DAI, Q., “Diagnosing error in object detectors,” in *ECCV*, 2012. 10
- [190] HOIEM, D., CHODPATHUMWAN, Y., and DAI, Q., “Diagnosing error in object detectors,” in *ECCV*, 2012.
- [191] HOVY, E., “Pragmatics and natural language generation,” *Artificial Intelligence*, 1990.
- [192] HU, R., ANDREAS, J., ROHRBACH, M., DARRELL, T., and SAENKO, K., “Learning to reason: End-to-end module networks for visual question answering,” 2017.
- [193] HUANG, T. K., FERRARO, F., MOSTAFAZADEH, N., MISRA, I., AGRAWAL, A., DEVLIN, J., GIRSHICK, R. B., HE, X., KOHLI, P., BATRA, D., ZITNICK, C. L., PARIKH, D., VANDERWENDE, L., GALLEY, M., and MITCHELL, M., “Visual storytelling,” in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2016.
- [194] HWANG, S. and GRAUMAN, K., “Reading between the lines: Object localization using implicit cues from image tags,” vol. 34, no. 6, pp. 1145–1158, 2012.
- [195] HWANG, S. J., GRAUMAN, K., and SHA, F., “Analogy-preserving semantic embedding for visual object categorization,” 2013.
- [196] HWANG, S. and GRAUMAN, K., “Learning the relative importance of objects from tagged images for retrieval and cross-modal search,” vol. 100, no. 2, pp. 134–153, 2012.
- [197] ILIEVSKI, I., YAN, S., and FENG, J., “A focused dynamic attention model for visual question answering,” *CoRR*, vol. abs/1604.01485, 2016. 3, 44
- [198] ISOLA, P., XIAO, J., TORRALBA, A., and OLIVA, A., “What makes an image memorable?,” in *CVPR*, 2011.
- [199] ISOLA, P., PARIKH, D., TORRALBA, A., and OLIVA, A., “Understanding the intrinsic memorability of images,” 2011.

- [200] ITTI, L., KOCH, C., and NIEBUR, E., “A model of saliency-based visual attention for rapid scene analysis,” vol. 20, no. 11, 1998.
- [201] ITTI, L., KOCH, C., and NIEBUR, E., “A model of saliency-based visual attention for rapid scene analysis,” vol. 20, no. 11, pp. 1254–1259, 1998.
- [202] JAIN, P., KULIS, B., DHILLON, I., and GRAUMAN, K., “Online metric learning and fast similarity search,” 2008.
- [203] JAIN, S., CHEN, Y., and PARKES, D. C., “Designing incentives for online question and answer forums,” in *Games and Economic Behavior (GEB)*, 2012.
- [204] JAMMALAMADAKA, N., ZISSERMAN, A., EICHNER, M., FERRARI, V., and JAWAHAR, C. V., “Has my algorithm succeeded? an evaluator for human pose estimators,” in *ECCV*, 2012.
- [205] JAS, M. and PARIKH, D., “Image Specificity,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [206] JAYARAMAN, D., SHA, F., and GRAUMAN, K., “Decorrelating semantic visual attributes by resisting the urge to share,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1629–1636, 2014. 11
- [207] JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., and DARRELL, T., “Caffe: Convolutional architecture for fast feature embedding,” *arXiv preprint arXiv:1408.5093*, 2014. 101
- [208] JIANG, A., WANG, F., PORIKLI, F., and LI, Y., “Compositional memory for visual question answering,” *CoRR*, vol. abs/1511.05676, 2015. 3, 44
- [209] JIANG, H., “Confidence measures for speech recognition: A survey,” *Speech Communication*, 2005.
- [210] JOACHIMS, T., “Optimizing search engines using clickthrough data,” 2002.
- [211] JOHNSON, J., HARIHARAN, B., VAN DER MAATEN, L., FEI-FEI, L., ZITNICK, C. L., and GIRSHICK, R., “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” 2017. 4, 12, 54, 55
- [212] JOHNSON, J., HARIHARAN, B., VAN DER MAATEN, L., FEI-FEI, L., ZITNICK, C. L., and GIRSHICK, R., “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1988–1997, IEEE, 2017.
- [213] JOHNSON, J., HARIHARAN, B., VAN DER MAATEN, L., HOFFMAN, J., FEI-FEI, L., ZITNICK, C. L., and GIRSHICK, R., “Inferring and executing programs for visual reasoning,” 2017.

- [214] JOHNSON, J., HARIHARAN, B., VAN DER MAATEN, L., HOFFMAN, J., FEI-FEI, L., ZITNICK, C. L., and GIRSHICK, R. B., “Inferring and executing programs for visual reasoning,” 2017.
- [215] JUDD, T., EHINGER, K., DURAND, F., and TORRALBA, A., “Learning to predict where humans look,” 2009.
- [216] KAFLE, K. and CHRISTOPHER, K., “An analysis of visual question answering algorithms,” in *ICCV*, 2017. 11
- [217] KAFLE, K. and KANAN, C., “Answer-type prediction for visual question answering,” in *CVPR*, 2016. 3, 44
- [218] KAFLE, K. and KANAN, C., “An analysis of visual question answering algorithms,” in *ICCV*, 2017. 74, 77
- [219] KANADE, T., “Recovery of the three-dimensional shape of an object from a single view,” *Artificial Intelligence*, vol. 17, no. 409–460, 1981.
- [220] KANEVA, B., TORRALBA, A., and FREEMAN, W. T., “Evaluation of image features using a photorealistic virtual world,” in *ICCV*, IEEE, 2011.
- [221] KAPPES, J. H., ANDRES, B., HAMPRECHT, F. A., SCHNÖRR, C., NOWOZIN, S., BATRA, D., KIM, S., KAUSLER, B. X., LELLMANN, J., KOMODAKIS, N., and ROTHER, C., “A comparative study of modern inference techniques for discrete energy minimization problems,” in *CVPR*, 2013.
- [222] KARPATHY, A. and FEI-FEI, L., “Deep Visual-Semantic Alignments for Generating Image Descriptions,” in *CVPR*, 2015. 9, 14
- [223] KARPATHY, A., JOHNSON, J., and LI, F., “Visualizing and understanding recurrent networks,” in *ICLR Workshop*, 2016. 10
- [224] KAWABATA, N., “Depth perception in simple line drawings,” *Perceptual Motor Skills*, 1997.
- [225] KAZEMZADEH, S., ORDONEZ, V., MATTEN, M., and BERG, T. L., “Refer-ItGame: Referring to Objects in Photographs of Natural Scenes,” in *EMNLP*, 2014. 10
- [226] KEKALAINEN, J. and JARVELIN, K., “Cumulated gain-based evaluation of IR techniques,” *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.
- [227] KIM, J.-H., LEE, S.-W., KWAK, D.-H., HEO, M.-O., KIM, J., HA, J.-W., and ZHANG, B.-T., “Multimodal residual learning for visual QA,” in *NIPS*, 2016. 3, 44
- [228] KINGMA, D. P. and BA, J., “Adam: A method for stochastic optimization,” 2014.

- [229] KIROS, R., SALAKHUTDINOV, R., and ZEMEL, R. S., “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models,” *TACL*, 2015. 9, 14
- [230] KIROS, R., ZHU, Y., SALAKHUTDINOV, R., ZEMEL, R. S., TORRALBA, A., URTASUN, R., and FIDLER, S., “Skip-thought vectors,” *arXiv preprint arXiv:1506.06726*, 2015. 100
- [231] KLEIN, D. and MANNING, C. D., “Accurate unlexicalized parsing,” 2003.
- [232] KOENDERINK, J., VAN DOORN, A., KAPPERS, A., and TODD, J., “Ambiguity and the ‘mental eye’ in pictorial relief,” *Perception*, 2001.
- [233] KOH, K., KIM, S.-J., and BOYD, S., “An interior-point method for large-scale l_1 -regularized logistic regression,” *J. Mach. Learn. Res.*, 2007.
- [234] KOHLI, P., KUMAR, M. P., and TORR, P. H. S., “ p^3 & beyond: Solving energies with higher order cliques,” in *CVPR*, 2007.
- [235] KONG, C., LIN, D., BANSAL, M., URTASUN, R., and FIDLER, S., “What Are You Talking About? Text-to-Image Coreference,” in *CVPR*, 2014. 10
- [236] KONG, C., LIN, D., BANSAL, M., URTASUN, R., and FIDLER, S., “What are you talking about? text-to-image coreference,” in *CVPR*, 2014.
- [237] KOSARAJU, D., SIMARD, C., and STEPHENSON, C., “Addressing core equity issues in K-12 computer science education: Identifying barriers and sharing strategies,” *Computer Science Teachers Association, Anita Borg Institute and University of Arizona*, 2009.
- [238] KOVASHKA, A., VIJAYANARASIMHAN, S., and GRAUMAN, K., “Actively selecting annotations among objects and attributes,” 2011.
- [239] KOVASHKA, A. and GRAUMAN, K., “Attribute adaptation for personalized image search,” in *ICCV*, 2013.
- [240] KOVASHKA, A., PARIKH, D., and GRAUMAN, K., “WhittleSearch: Image search with relative attribute feedback,” 2012.
- [241] KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANTIDIS, Y., LI, L.-J., SHAMMA, D. A., and OTHERS, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *arXiv preprint arXiv:1602.07332*, 2016.
- [242] KRIZHEVSKY, A., SUTSKEVER, I., and HINTON, G., “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.
- [243] KRIZHEVSKY, A., SUTSKEVER, I., and HINTON, G. E., “ImageNet Classification with Deep Convolutional Neural Networks,” in *NIPS*, 2012. 9, 16

- [244] KUENZI, J. J., “Science, Technology, Engineering, and Mathematics (STEM) Education: Background, Federal Policy, and Legislative Action.” Congressional Research Service (CRS) report for Congress <http://www.fas.org/sgp/crs/misc/RL33434.pdf>, 2008.
- [245] KUKAR, M., “Estimating confidence values of individual predictions by their typicalness and reliability,” in *ECAI*, 2004.
- [246] KULESZA, A. and TASKAR, B., “Determinantal point processes for machine learning,” *Foundations and Trends in Machine Learning*, vol. 5, no. 2–3, 2012.
- [247] KULKARNI, G., PREMRAJ, V., SAGNIK DHAR AND, S. L., CHOI, Y., BERG, A. C., and BERG, T. L., “Baby Talk: Understanding and Generating Simple Image Descriptions,” in *CVPR*, 2011. 9
- [248] KULKARNI, T. D., KOHLI, P., TENENBAUM, J. B., and MANSINGHKA, V., “Picture: A probabilistic programming language for scene perception,” pp. 4390–4399, 2015.
- [249] KULKARNI, T. D., WHITNEY, W. F., KOHLI, P., and TENENBAUM, J., “Deep convolutional inverse graphics network,” pp. 2539–2547, 2015.
- [250] KUMAR, N., BELHUMEUR, P., and NAYAR, S., “Facetracer: A search engine for large collections of images with faces,” 2010.
- [251] KUMAR, N., BERG, A., BELHUMEUR, P., and NAYAR, S., “Attribute and simile classifiers for face verification,” 2009.
- [252] KUMAR, N., BELHUMEUR, P. N., BISWAS, A., JACOBS, D. W., KRESS, W. J., LOPEZ, I., and SOARES, J. V. B., “Leafsnap: A computer vision system for automatic plant species identification,” 2012.
- [253] KURITA, T. and KATO, T., “Learning of personal visual impression for image database systems,” 1993.
- [254] LALONDE, J. F., EFROS, A. A., and NARASIMHAN, S. G., “Detecting ground shadows in outdoor consumer photographs,” in *ECCV*, 2010.
- [255] LAMPERT, C., NICKISCH, H., and HARMELING, S., “Learning to detect unseen object classes by between-class attribute transfer,” 2009.
- [256] LAMPERT, C. H., NICKISCH, H., and HARMELING, S., “Learning to detect unseen object classes by between-class attribute transfer,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 951–958, IEEE, 2009. 11
- [257] LAMPLE, G., ZEGHIDOUR, N., USUNIER, N., BORDES, A., DENOYER, L., and RANZATO, M., “Fader networks: Manipulating images by sliding attributes,” 2017. 13

- [258] LANCKRIET, G., CRISTIANINI, N., BARTLETT, P., GHAOUI, L. E., and JORDAN, M., “Learning the kernel matrix with semidefinite programming,” vol. 5, pp. 27–72, 2004.
- [259] LAZEBNIK, S., SCHMID, C., and PONCE, J., “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” 2006.
- [260] LEE, D., HEBERT, M., and KANADE, T., “Geometric reasoning for single image structure recovery,” 2009.
- [261] LENAT, D. B. and GUHA, R. V., *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., 1989. 16
- [262] LEUNG, T. and MALIK, J., “Contour continuity in region based image segmentation,” in *ECCV*, 1998.
- [263] LI, B., CHANG, E., and LI, C.-S., “Learning image query concepts via intelligent sampling,” 2001.
- [264] LI, C., PARIKH, D., and CHEN, T., “Extracting adaptive contextual cues from unlabeled regions,” 2011.
- [265] LI, L.-J., SU, H., XING, E. P., and FEI-FEI, L., “Object Bank: A high-level image representation for scene classification and semantic feature sparsification,” 2010.
- [266] LI, Y. and HUTTENLOCHER, D. P., “Sparse long-range random field and its application to image denoising,” 2008.
- [267] LIN, C.-Y., “Rouge: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (MARIE-FRANCINE MOENS, S. S., ed.), (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
- [268] LIN, C. H., MAUSAM, and WELD, D. S., “Crowdsourcing control: Moving beyond multiple choice,” in *UAI*, 2012.
- [269] LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., and ZITNICK, C. L., “Microsoft COCO: Common Objects in Context,” in *ECCV*, 2014. 2, 16, 17, 18, 19, 23
- [270] LIN, W.-H. and HAUPTMANN, A., “Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts,” 2006.
- [271] LIN, X. and PARIKH, D., “Don’t Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks,” in *CVPR*, 2015. 16
- [272] LIN, X. and PARIKH, D., “Don’t just listen, use your imagination: Leveraging visual common sense for non-visual tasks,” in *CVPR*, 2015. 9

- [273] LIU, H. and SINGH, P., “ConceptNet — A Practical Commonsense Reasoning Tool-Kit,” *BT Technology Journal*, 2004. 16
- [274] LIU, L. and WANG, L., “What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection,” in *CVPR*, 2012.
- [275] LIU, T., YUAN, Z., SUN, J., WANG, J., ZHENG, N., TANG, X., and SHUM, H., “Learning to detect a salient object,” vol. 33, no. 2, 2011.
- [276] LIU, Z. and KERSTEN, D., “2d observers for human 3d object recognition?,” *Vision Research*, 1998.
- [277] LOUPPE, G., KAGAN, M., and CRANMER, K., “Learning to pivot with adversarial networks,” pp. 982–991, 2017. 13
- [278] LOWE, D. G., “Distinctive image features from scale-invariant keypoints,” vol. 60, no. 2, pp. 91–110, 2004.
- [279] LU, J., LIN, X., BATRA, D., and PARIKH, D., “Deeper lstm and normalized cnn visual question answering model.” https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 44, 45, 55, 58, 59
- [280] LU, J., YANG, J., BATRA, D., and PARIKH, D., “Hierarchical question-image co-attention for visual question answering,” in *NIPS*, 2016. 3, 44, 45
- [281] LU, J., YANG, J., BATRA, D., and PARIKH, D., “Hierarchical question-image co-attention for visual question answering,” 2016.
- [282] MA, W. and MANJUNATH, B., “NeTra: a toolbox for navigating large image databases,” in *ICIP*, 1997.
- [283] MAHAJAN, D., SELAMANICKAM, S., and NAIR, V., “A joint learning framework for attribute models and object descriptions,” 2011.
- [284] MALINOWSKI, M. and DOERSCH, C., “The visual qa devil in the details: The impact of early fusion and batch norm on clevr,” *arXiv preprint arXiv:1809.04482*, 2018.
- [285] MALINOWSKI, M., DOERSCH, C., SANTORO, A., and BATTAGLIA, P., “Learning visual question answering by bootstrapping hard attention,” pp. 3–20, 2018.
- [286] MALINOWSKI, M. and FRITZ, M., “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input,” in *NIPS*, 2014. 8, 15
- [287] MALINOWSKI, M. and FRITZ, M., “A multi-world approach to question answering about real-world scenes based on uncertain input,” in *Advances in Neural Information Processing Systems*, pp. 1682–1690, 2014.

- [288] MALINOWSKI, M. and FRITZ, M., “Towards a visual turing challenge,” *arXiv preprint arXiv:1410.8027*, 2014.
- [289] MALINOWSKI, M., ROHRBACH, M., and FRITZ, M., “Ask your neurons: A neural-based approach to answering questions about images,” in *ICCV*, 2015. 8
- [290] MAO, J., XU, W., YANG, Y., WANG, J., and YUILLE, A. L., “Explain Images with Multimodal Recurrent Neural Networks,” *CoRR*, vol. abs/1410.1090, 2014. 9, 14
- [291] MARGOLIS, J. and FISHER, A., “Geek mythology and attracting undergraduate women to computer science,” in *Impacting Change Through Collaboration, Proceedings of the Joint National Conference of the Women in Engineering Program Advocates Network and the National Association of Minority Engineering Program Administrators*, 1997.
- [292] MARIN, J., VAZQUEZ, D., GERONIMO, D., and LOPEZ, A., “Learning appearance in virtual scenarios for pedestrian detection,” in *CVPR*, 2007.
- [293] MARR, D., *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, 1982.
- [294] MARTIN, D., FOWLKES, C., TAL, D., and MALIK, J., “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *ICCV*, 2001.
- [295] MARTIN, D. R., FOWLKES, C. C., and MALIK, J., “Learning to detect natural image boundaries using local brightness, color, and texture cues,” *PAMI*, vol. 26, pp. 530–549, 2004.
- [296] MASCHARKA, D., TRAN, P., SOKLASKI, R., and MAJUMDAR, A., “Transparency by design: Closing the gap between performance and interpretability in visual reasoning,” 2018.
- [297] MCDERMOTT, J., “Psychophysics with junctions in real images,” *Perception*, vol. 33, pp. 1101–1127, 2004.
- [298] MELTZER, T., YANOVER, C., and WEISS, Y., “Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation,” in *ICCV*, pp. 428–435, 2005.
- [299] METZ, C., “Facebook AI Can Caption Photos for the Blind on Its Own.” <http://www.wired.com/2015/10/facebook-artificial-intelligence-describes-photo-captions-for-blind-people/>, October 2015.
- [300] MIKOLAJCZYK, K. and SCHMID, C., “A performance evaluation of local descriptors,” *PAMI*, 2005.

- [301] MIKOLOV, T., CHEN, K., CORRADO, G., and DEAN, J., “Efficient estimation of word representations in vector space,” in *ICLR*, 2013. 47
- [302] MIKOLOV, T., KARAFIÁT, M., BURGET, L., CERNOCKÝ, J., and KHUNDANPUR, S., “Recurrent neural network based language model,” in *INTER-SPEECH*, pp. 1045–1048, 2010.
- [303] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., and DEAN, J., “Distributed Representations of Words and Phrases and their Compositionality,” in *NIPS*, 2013. 21, 31, 101
- [304] MILLER, G. A., “WordNet: A lexical database for English,” *Communications of the ACM*, vol. 38, pp. 39–41, 1995.
- [305] MILLER, G. A., “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [306] MINSKY, M., *Society of Mind*. Simon & Schuster, 1988.
- [307] MIRZA, M. and OSINDERO, S., “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014. 13
- [308] MISRA, I., GUPTA, A., and HEBERT, M., “From red wine to red tomato: Composition with context,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1160–1169, 2017.
- [309] MITCHELL, M., DODGE, J., GOYAL, A., YAMAGUCHI, K., STRATOS, K., HAN, X., MENSCH, A., BERG, A., BERG, T. L., and DAUME III, H., “Midge: Generating Image Descriptions From Computer Vision Detections,” in *ACL*, 2012. 9
- [310] MITCHELL, M., VAN DEEMTER, K., and REITER, E., “Two approaches for generating size modifiers,” in *Proceedings of the European Workshop on Natural Language Generation*, 2011.
- [311] MITCHELL, M., VAN DEEMTER, K., and REITER, E., “Attributes in visual reference,” in *PRE-CogSci*, 2013. 94
- [312] MITCHELL, M., VAN DEEMTER, K., and REITER, E., “Generating Expressions that Refer to Visible Objects,” in *HLT-NAACL*, 2013. 10
- [313] MNIH, V., BADIA, A. P., MIRZA, M., GRAVES, A., LILLICRAP, T., HARLEY, T., SILVER, D., and KAVUKCUOGLU, K., “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*, pp. 1928–1937, 2016.
- [314] MOSTAFAZADEH, N., BROCKETT, C., DOLAN, B., GALLEY, M., GAO, J., SPITHOURAKIS, G. P., and VANDERWENDE, L., “Image-grounded conversations: Multimodal context for natural question and response generation,” *CoRR*, vol. abs/1701.08251, 2017.

- [315] MOTTAGHI, R., FIDLER, S., YAO, J., URTASUN, R., and PARIKH, D., “Analyzing semantic segmentation using hybrid human-machine CRFs,” 2013.
- [316] MUHLBAIER, M., TOPALIS, A., and POLIKAR, R., “Ensemble confidence estimates posterior probability,” in *International Conference on Multiple Classifier Systems*, 2005.
- [317] MUNOZ, D., BAGNELL, J. A. D., and HEBERT, M., “Stacked hierarchical labeling,” 2010.
- [318] NAIR, V. and HINTON, G. E., “Inferring motor programs from images of handwritten digits,” pp. 515–522, 2006.
- [319] NAIR, V., SUSSKIND, J., and HINTON, G. E., “Analysis-by-synthesis by learning to invert generative black boxes,” pp. 971–981, Springer, 2008.
- [320] NAPHADE, M., SMITH, J., TESIC, J., CHANG, S., HSU, W., KENNEDY, L., HAUPTMANN, A., and CURTIS, J., “Large-scale concept ontology for multimedia,” *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.
- [321] NOH, H. and HAN, B., “Training recurrent answering units with joint loss minimization for vqa,” *CoRR*, vol. abs/1606.03647, 2016. 3, 44
- [322] OATLEY, K. and YUILL, N., “Perception of personal and interpersonal action in a cartoon film,” *British J. of Social Psychology*, vol. 24, no. 2, 2011.
- [323] OHTA, Y., *Knowledge-Based Interpretation Of Outdoor Natural Color Scenes*. Pitman, 1985.
- [324] OHTA, Y., KANADE, T., and SAKAI, T., “An analysis system for scenes containing objects with substructures,” 1978.
- [325] OLIVA, A. and SCHYNS, P. G., “Diagnostic colors mediate scene recognition,” *Cognitive Psychology*, 2000.
- [326] OLIVA, A. and TORRALBA, A., “Modeling the shape of the scene: A holistic representation of the spatial envelope,” vol. 42, pp. 145–175, 2001.
- [327] OLIVA, A., TORRALBA, A., and OTHERS, “The role of context in object recognition,” *Trends in cognitive sciences*, vol. 11, no. 12, 2007.
- [328] ORDONEZ, V., KULKARNI, G., and BERG, T., “Im2text: Describing images using 1 million captioned photographs,” 2011.
- [329] PAPINENI, K., ROUKOS, S., WARD, T., and ZHU, W.-J., “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02*, (Stroudsburg, PA, USA), pp. 311–318, Association for Computational Linguistics, 2002.

- [330] PARIKH, D., “Recognizing jumbled images: The role of local and global information in image classification,” 2011.
- [331] PARIKH, D. and GRAUMAN, K., “Interactively building a discriminative vocabulary of nameable attributes,” 2011.
- [332] PARIKH, D. and GRAUMAN, K., “Relative attributes,” 2011.
- [333] PARIKH, D. and ZITNICK, C. L., “The role of features, algorithms and data in visual recognition,” 2010.
- [334] PARIKH, D. and ZITNICK, C. L., “Finding the weakest link in person detectors,” 2011.
- [335] PARIKH, D. and ZITNICK, C. L., “Human-debugging of machines,” in *Second Workshop on Computational Social Science and the Wisdom of Crowds, Neural Information Processing Systems (NIPS)*, 2011.
- [336] PARIKH, D., ZITNICK, C. L., and CHEN, T., “From appearance to context-based recognition: Dense labeling in small images,” 2008.
- [337] PARIKH, D., ZITNICK, C. L., and CHEN, T., “Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition,” vol. 34, no. 10, pp. 1978–1991, 2012.
- [338] PARIKH, D. and CHEN, T., “Hierarchical semantics of objects (hsos),” in *ICCV*, 2007.
- [339] PARIKH, D. and CHEN, T., “Unsupervised modeling of objects and their hierarchical contextual interactions,” *EURASIP Journal on Image and Video Processing, Special Issue on Patches in Vision*, 2008.
- [340] PARIKH, D. and GRAUMAN, K., “Implied feedback: Learning nuances of user behavior in image search,” in *ICCV*, 2013.
- [341] PARIKH, D., ZITNICK, C. L., and CHEN, T., “Unsupervised learning of hierarchical spatial structures in images,” in *CVPR*, 2000.
- [342] PARIKH, D., ZITNICK, C. L., and CHEN, T., “Determining patch saliency using low-level context,” in *ECCV*, 2008.
- [343] PARKASH, A. and PARIKH, D., “Attributes for classifier feedback,” 2012.
- [344] PATTERSON, G. and HAYS, J., “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” 2012.
- [345] PEARL, J., “Reverend bayes on inference engines: A distributed hierarchical approach,” in *American Association of Artificial Intelligence National Conference on AI*, 1982.

- [346] PENNINGTON, J., SOCHER, R., and MANNING, C. D., “Glove: Global vectors for word representation,” in *EMNLP*, 2014. 64
- [347] PEREZ, E., DE VRIES, H., STRUB, F., DUMOULIN, V., and COURVILLE, A., “Learning visual reasoning without strong priors,” 2017.
- [348] PEREZ, E., STRUB, F., DE VRIES, H., DUMOULIN, V., and COURVILLE, A., “Film: Visual reasoning with a general conditioning layer,” 2018.
- [349] PETERSON, M., “Object recognition processes can and do operate before figure-ground organization,” *Current Directions in Psychological Science*, vol. 3, 1994.
- [350] PETRIE, H., HARRISON, C., and DEV, S., “Describing Images on the Web: a Survey of Current Practice and Prospects for the Future,” in *Proceedings of Human Computer Interaction International (HCII)*, July 2005.
- [351] PIRSIAVASH, H., VONDRICK, C., and TORRALBA, A., “Inferring the why in images,” *CoRR*, vol. abs/1406.5472, 2014.
- [352] PRIVITERA, C. M. and STARK., L. W., “Algorithms for defining visual regions-of-interest: Comparison with eye fixations,” *PAMI*, 2000.
- [353] PRIVITERA, C. and STARK, L., “Algorithms for defining visual regions-of-interest: Comparison with eye fixations,” vol. 22, no. 9, 2000.
- [354] QUIRK, C., CHOUDHURY, P., GAO, J., SUZUKI, H., TOUTANOVA, K., GAMON, M., YIH, W.-T., VANDERWENDE, L., and CHERRY, C., “Msr splat, a language analysis toolkit,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstration Session*, pp. 21–24, Association for Computational Linguistics, 2012.
- [355] RABINOVICH, A., VEDALDI, A., GALLEGUILLOS, C., WIEWIORA, E., and BELONGIE, S., “Objects in context,” 2007.
- [356] RADFORD, A., METZ, L., and CHINTALA, S., “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015. 13
- [357] RAGHAVAN, H., MADANI, O., and JONES, R., “Interactive feature selection,” 2005.
- [358] RAKOTOMAMONJY, A., BACH, F., CANU, S., and GRANDVALET, Y., “More efficiency in multiple kernel learning,” 2007.
- [359] RAMAKRISHNAN, S., AGRAWAL, A., and LEE, S., “Overcoming language priors in visual question answering with adversarial regularization,” in *NIPS*, 2018. 5

- [360] RAMAKRISHNAN, S. K., PAL, A., SHARMA, G., and MITTAL, A., “An empirical evaluation of visual question answering for novel objects,” *arXiv preprint arXiv:1704.02516*, 2017. 12
- [361] RAMANATHAN, V., JOULIN, A., LIANG, P., and FEI-FEI, L., “Linking People with “Their” Names using Coreference Resolution,” in *ECCV*, 2014. 10
- [362] RANDEN, T. and HUSOEY, J., “Filtering for texture classification: A comparative study,” *PAMI*, 1999.
- [363] RASHTCHIAN, C., YOUNG, P., HODOSH, M., and HOCKENMAIER, J., “Collecting image annotations using amazon’s mechanical turk,” in *NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s MT*, 2010.
- [364] RASHTCHIAN, C., YOUNG, P., HODOSH, M., and HOCKENMAIER, J., “Collecting Image Annotations Using Amazon’s Mechanical Turk,” in *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010.
- [365] RASHTCHIAN, C., YOUNG, P., HODOSH, M., and HOCKENMAIER, J., “Collecting image annotations using amazon’s mechanical turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010.
- [366] RASIWASIA, N., MORENO, P., and VASCONCELOS, N., “Bridging the gap: Query by semantic example,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.
- [367] RASTEGARI, M., FARHADI, A., and FORSYTH, D., “Attribute discovery via predictable discriminative binary codes,” 2012.
- [368] REED, S. E., AKATA, Z., YAN, X., LOGESWARAN, L., SCHIELE, B., and LEE, H., “Generative adversarial text to image synthesis,” *CoRR*, vol. abs/1605.05396, 2016.
- [369] REN, M., KIROS, R., and ZEMEL, R., “Exploring models and data for image question answering,” in *NIPS*, 2015. 9
- [370] REN, M., KIROS, R., and ZEMEL, R., “Exploring models and data for image question answering,” in *Advances in Neural Information Processing Systems*, pp. 2953–2961, 2015.
- [371] REN, S., HE, K., GIRSHICK, R., and SUN, J., “Faster r-cnn: Towards real-time object detection with region proposal networks,” 2015. 80
- [372] RICHARDSON, M., BURGESS, C. J., and RENSHAW, E., “MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text,” in *EMNLP*, 2013. 9, 16

- [373] RIVEST, J. and CABANAGH, P., “Localizing contours defined by more than one attribute,” *Vision Research*, 1996.
- [374] ROHRBACH, M., QIU, W., TITOV, I., THATER, S., PINKAL, M., and SCHIELE, B., “Translating Video Content to Natural Language Descriptions,” in *ICCV*, 2013. 9
- [375] ROTHER, C., KOHLI, P., FENG, W., and JIA, J., “Minimizing sparse higher order energy functions of discrete variables,” in *CVPR*, 2009.
- [376] RUI, Y., HUANG, T. S., ORTEGA, M., and MEHROTRA, S., “Relevance feedback: A power tool for interactive content-based image retrieval,” *IEEE Transactions on Circuits and Video Technology*, vol. 8, no. 5, pp. 644–655, 1998.
- [377] RUSSELL, B., TORRALBA, A., MURPHY, K., and FREEMAN, W., “LabelMe: A database and web-based tool for image annotation,” vol. 77, no. 1-3, pp. 157–173, 2008.
- [378] SADEGHI, F., KUMAR DIVVALA, S. K., and FARHADI, A., “Viske: Visual knowledge extraction and question answering by visual verification of relation phrases,” in *CVPR*, 2015. 9
- [379] SADEGHI, M. and FARHADI, A., “Recognition using visual phrases,” 2011.
- [380] SADOVNIK, A., GALLAGHER, A. C., PARIKH, D., and CHEN, T., “Spoken attributes: Mixing binary and relative attributes to say the right thing,” in *ICCV*, 2013.
- [381] SAITO, K., SHIN, A., USHIKU, Y., and HARADA, T., “Dualnet: Domain-invariant network for visual question answering,” *CoRR*, vol. abs/1606.06108, 2016. 3, 44
- [382] SALEH, B., FARHADI, A., and ELGAMMAL, A., “Object-centric anomaly detection by attribute-based reasoning,” 2013.
- [383] SALIMANS, T., GOODFELLOW, I. J., ZAREMBA, W., CHEUNG, V., RADFORD, A., and CHEN, X., “Improved techniques for training gans,” *CoRR*, vol. abs/1606.03498, 2016.
- [384] SANTORO, A., RAPOSO, D., BARRETT, D. G., MALINOWSKI, M., PASCANU, R., BATTAGLIA, P., and LILLICRAP, T., “A simple neural network module for relational reasoning,” pp. 4967–4976, 2017.
- [385] SARMA, A. and PALMER, D. D., “Context-based speech recognition error detection and correction,” in *NAACL (Short papers)*, 2004.
- [386] SATKIN, S., LIN, J., and HEBERT, M., “Data-driven scene understanding from 3D models,” 2012.

- [387] SAUL, L. K. and ROWEIS, S. T., “An introduction to locally linear embedding.”
- [388] SAXENA, A., DRIEMEYER, J., and NG, A. Y., “Robotic grasping of novel objects using vision,” *IJRR*, 2008.
- [389] SAXENA, A., SUN, M., and NG, A. Y., “Make3d: Learning 3d scene structure from a single still image,” *PAMI*, 2009.
- [390] SCHELS, J., LIEBELT, J., SCHERTLER, K., and LIENHART, R., “Building a semantic part-based object class detector from synthetic 3d models,” in *ICME*, 2011.
- [391] SCHMID, C., MOHR, R., and BAUCKHAGE, C., “Evaluation of interest point detectors,” *IJCV*, 2000.
- [392] SCHULTZ, M. and JOACHIMS, T., “Learning a distance metric from relative comparisons,” 2003.
- [393] SCHWAB, K., “The Global Competitiveness Report 2010-2011.” World Economic Forum. <http://reports.weforum.org/global-competitiveness-2011-2012>, 2011.
- [394] SECURICS, “Mughunt2.” <http://mughunt.securics.com/>.
- [395] SETTLES, B., “Active learning literature survey,” tech. rep., 2010.
- [396] SETTLES, B., *Active Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning)*. Morgan and Claypool Publishers, 2012.
- [397] SHAKHNAROVICH, G., “Learning task-specific similarity,” in *Ph.D. Thesis, MIT*, 2006.
- [398] SHECHTMAN, E. and IRANI, M., “Matching local self-similarities across images and videos,” 2007.
- [399] SHI, J. and MALIK, J., “Normalized cuts and image segmentation,” *PAMI*, 2000.
- [400] SHIH, K. J., SINGH, S., and HOIEM, D., “Where to look: Focus regions for visual question answering,” in *CVPR*, 2016. 3, 44
- [401] SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., and BLAKE, A., “Real-time human pose recognition in parts from single depth images,” in *CVPR*, 2011.
- [402] SHOTTON, J., WINN, J., ROTHER, C., and CRIMINISI, A., “Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation,” in *ECCV*, 2006.

- [403] SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., and BLAKE, A., “Real-time human pose recognition in parts from single depth images,” *CVPR*, 2011.
- [404] SHRIVASTAVA, A., SINGH, S., and GUPTA, A., “Constrained semi-supervised learning using attributes and comparative attributes,” 2012.
- [405] SIDDIQUIE, B. and GUPTA, A., “Beyond active noun tagging: Modeling contextual interactions for multi-class active learning,” 2010.
- [406] SIDDIQUIE, B., FERIS, R. S., and DAVIS, L. S., “Image ranking and retrieval based on multi-attribute queries,” 2011.
- [407] SILBERMAN, N., HOIEM, D., KOHLI, P., and FERGUS, R., “Indoor segmentation and support inference from rgb-d images,” in *ECCV*, 2012.
- [408] SIMONYAN, K. and ZISSERMAN, A., “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. xii, 32, 37
- [409] SIMONYAN, K. and ZISSERMAN, A., “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.
- [410] SIMONYAN, K. and ZISSERMAN, A., “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. 63, 125, 127
- [411] SIMONYAN, K. and ZISSERMAN, A., “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. 80
- [412] SINGH, A., NATARAJAN, V., SHAH, M., JIANG, Y., CHEN, X., BATRA, D., PARIKH, D., and ROHRBACH, M., “Towards vqa models that can read,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 89
- [413] SINGH, P., LIN, T., MUELLER, E. T., LIM, G., PERKINS, T., and ZHU, W. L., “Open mind common sense: Knowledge acquisition from the general public,” in *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE*, Springer-Verlag, 2002.
- [414] SMITH, J., NAPHADE, M., and NATSEV, A., “Multimedia semantic indexing using model vectors,” 2003.
- [415] SNAVELY, N., SEITZ, S. M., and SZELISKI, R., “Photo tourism: Exploring photo collections in 3d,” in *SIGGRAPH*, 2006.
- [416] SOCHER, R., BAUER, J., MANNING, C. D., and NG, A. Y., “Parsing with compositional vector grammars,” 2013.
- [417] SONNENBURG, S., RTSCH, G., SCHFER, C., and SCHLKOPF, B., “Large scale multiple kernel learning,” vol. 7, pp. 1531–1565, 2006.

- [418] SONTAG, D., MELTZER, T., GLOBERSON, A., JAAKKOLA, T., and WEISS, Y., “Tightening lp relaxations for map using message passing,” in *UAI*, 2008.
- [419] SPAIN, M. and PERONA, P., “Measuring and predicting object importance,” vol. 91, no. 1, pp. 59–76, 2011.
- [420] STACK, J., “Automation for underwater mine recognition: Current trends & future strategy,” in *Proceedings of SPIE Defense & Security*, 2011.
- [421] SUTSKEVER, I., VINYALS, O., and LE, Q., “Sequence to Sequence Learning with Neural Networks,” in *Neural Information Processing Systems (NIPS)*, 2014.
- [422] SZELISKI, R., ZABIH, R., SCHARSTEIN, D., VEKSLER, O., KOLMOGOROV, V., AGARWALA, A., TAPPEN, M., and ROTHER, C., “A comparative study of energy minimization methods for markov random fields with smoothness-based priors,” *PAMI*, vol. 30, no. 6, pp. 1068–1080, 2008.
- [423] SZUMMER, M., KOHLI, P., and HOIEM, D., “Learning crfs using graph cuts,” in *ECCV*, 2008.
- [424] TAMUZ, O., LIU, C., BELONGIE, S., SHAMIR, O., and KALAI, A. T., “Adaptively learning the crowd kernel,” 2011.
- [425] TARR, M. J. and PINKER, S., “When does human object recognition use a viewer-centered reference frame?,” *Psychological Science*, 1990.
- [426] TAYLOR, G. R., CHOSAK, A. J., and BREWER, P. C., “Ovvv: Using virtual worlds to design and evaluate surveillance systems,” in *CVPR*, 2007.
- [427] TENENBAUM, J. B., DE SILVA, V., and LANGFORD, J. C., “A global geometric framework for nonlinear dimensionality reduction,” *Science Magazine*, December 2000.
- [428] TENEY, D. and HENGEL, A. v. D., “Zero-shot visual question answering,” *arXiv preprint arXiv:1611.05546*, 2016. 12
- [429] TIEU, K. and VIOLA, P., “Boosting image retrieval,” 2000.
- [430] TODD, J. T., “The visual perception of 3d shape,” in *Trends in Cognitive Science*, 2004.
- [431] TONG, S. and CHANG, E., “Support Vector Machine active learning for image retrieval,” in *Proceedings of ACM Multimedia*, 2001.
- [432] TORRALBA, A. and EFROS, A., “Unbiased look at dataset bias,” 2011.
- [433] TORRALBA, A., FERGUS, R., and FREEMAN, W. T., “80 million tiny images: A large dataset for non-parametric object and scene recognition,” vol. 30, no. 11, pp. 1958–1970, 2008.

- [434] TORRALBA, A., MURPHY, K., and FREEMAN, W., “Contextual models for object detection using boosted random fields,” 2004.
- [435] TORRALBA, A. and EFROS, A. A., “Unbiased look at dataset bias,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1521–1528, IEEE, 2011.
- [436] TORRESANI, L., SZUMMER, M., and FITZGIBBON, A., “Efficient object category recognition using classemes,” 2010.
- [437] TOUTANOVA, K., KLEIN, D., MANNING, C. D., and SINGER, Y., “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *ACL*, 2003. 93
- [438] TSENG, P., CARMI, R., CAMERON, I., MUNOZ, D., and ITTI, L., “Quantifying center bias of observers in free viewing of dynamic natural scenes,” *Journal of Vision*, vol. 9, no. 7, 2009.
- [439] TU, K., MENG, M., LEE, M. W., CHOE, T. E., and ZHU, S. C., “Joint Video and Text Parsing for Understanding Events and Answering Queries,” *IEEE MultiMedia*, 2014. 8, 15
- [440] TURAKHIA, N. and PARIKH, D., “Attribute dominance: What pops out?,” in *ICCV*, 2013.
- [441] TURING, A. M., “Computing machinery and intelligence,” *Mind*, vol. 59, 1950. 1
- [442] TZENG, E., HOFFMAN, J., DARRELL, T., and SAENKO, K., “Adversarial discriminative domain adaptation,” 2017. 13
- [443] ULLMAN, S., VIDAL-NAQUET, M., SALI, E., and OTHERS, “Visual features of intermediate complexity and their use in classification,” *Nature neuroscience*, vol. 5, no. 7, 2002.
- [444] VAN DER MAATEN, L., “Matlab toolbox for dimensionality reduction (v0.8.1 - march 2013).” http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html.
- [445] VAQUERO, D. A., FERIS, R. S., TRAN, D., BROWN, L., HAMPAPUR, A., and TURK, M., “Attribute-based people search in surveillance environments,” in *WACV*, 2009.
- [446] VEDANTAM, R., ZITNICK, C. L., and PARIKH, D., “Collecting Image Description Datasets using Crowdsourcing,” *arXiv preprint arXiv:1411.3041*, 2014.
- [447] VEDANTAM, R., ZITNICK, C. L., and PARIKH, D., “CIDeR: Consensus-based Image Description Evaluation,” in *CVPR*, 2015. 14

- [448] VENDANTAM, R., LIN, X., BATRA, T., ZITNICK, C. L., and PARIKH, D., “Learning common sense through visual abstraction,” in *ICCV*, 2015. 9
- [449] VIJAYANARASIMHAN, S. and GRAUMAN, K., “Multi-level active prediction of useful image annotations for recognition,” 2008.
- [450] VIJAYANARASIMHAN, S. and GRAUMAN, K., “Large-scale live active learning: Training object detectors with crawled data and crowds,” 2011.
- [451] VINYALS, O., TOSHEV, A., BENGIO, S., and ERHAN, D., “Show and Tell: A Neural Image Caption Generator,” in *CVPR*, 2015. 9, 14
- [452] VON AHN, L. and DABBISH, L., “Labeling images with a computer game,” 2004.
- [453] VON AHN, L. and DABBISH, L., “Labeling images with a computer game,” 2004.
- [454] VONDRICK, C., KHOSLA, A., MALISIEWICZ, T., and TORRALBA, A., “Inverting and visualizing features for object detection,” *arXiv preprint arXiv:1212.2278*, 2012.
- [455] WAGNER, M., BASEVI, H., SHETTY, R., LI, W., MALINOWSKI, M., FRITZ, M., and LEONARDIS, A., “Answering visual what-if questions: From actions to predicted scene descriptions,” *arXiv preprint arXiv:1809.03707*, 2018.
- [456] WAINWRIGHT, M., JAAKKOLA, T., and WILLSKY, A., “Map estimation via agreement on (hyper)trees: Message-passing and linear programming approaches,” *IEEE Transactions on Information Theory*, 2002.
- [457] WALTHER, D. B., CHAI, B., CADDIGAN, E., BECK, D. M., and FEI-FEI, L., “Simple line drawings suffice for functional mri decoding of natural scene categories,” in *PNAS*, 2011.
- [458] WALTZ, D., “Generating semantic descriptions from drawings of scenes with shadows,” tech. rep., MIT, 1972.
- [459] WANG, G. and FORSYTH, D., “Joint learning of visual attributes, object classes and visual saliency,” 2009.
- [460] WANG, G., FORSYTH, D., and HOIEM, D., “Comparative object similarity for improved recognition with few or no examples,” 2010.
- [461] WANG, G., FORSYTH, D., and HOIEM, D., “Comparative object similarity for improved recognition with few or no examples,” 2010.
- [462] WANG, G. and FORSYTH, D., “Joint learning of visual attributes, object classes and visual saliency,” 2009.

- [463] WANG, J., MARKERT, K., and EVERINGHAM, M., “Learning models for object recognition from natural language descriptions,” 2009.
- [464] WANG, P., WU, Q., SHEN, C., HENGEL, A., and DICK, A., “Fvqa: Fact-based visual question answering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 2016.
- [465] WANG, P., WU, Q., SHEN, C., VAN DEN HENGEL, A., and DICK, A. R., “Explicit knowledge-based reasoning for visual question answering,” *CoRR*, vol. abs/1511.02570, 2015. 3, 44
- [466] WANG, R. and BHANU, B., “Learning models for predicting recognition performance,” in *ICCV*, 2005.
- [467] WANG, X., LIU, K., and TANG, X., “Query-specific visual semantic spaces for web image re-ranking,” 2011.
- [468] WANG, Y. and MORI, G., “A discriminative latent model of object classes and attributes,” 2010.
- [469] WEINBERGER, K., BLITZER, J., and SAUL, L., “Distance metric learning for large margin nearest neighbor classification,” 2006.
- [470] WEISGRAM, E. and BIGLER, R., “The role of attitudes and intervention in high school girls’ interest in computer science,” *Journal of Women and Minorities in Science and Engineering*, vol. 12, no. 4, pp. 325–336, 2006.
- [471] WELINDER, P., BRANSON, S., BELONGIE, S., and PERONA, P., “The multi-dimensional wisdom of crowds,” in *NIPS*, 2010.
- [472] WESTON, J., BENGIO, S., and USUNIER, N., “Large scale image annotation: learning to rank with joint word-image embeddings,” *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.
- [473] WESTON, J., BORDES, A., CHOPRA, S., and MIKOLOV, T., “Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks,” *CoRR*, vol. abs/1502.05698, 2015. 9
- [474] WESTON, J., BORDES, A., CHOPRA, S., and MIKOLOV, T., “Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks,” *CoRR*, vol. abs/1502.05698, 2015.
- [475] WESTON, J., CHOPRA, S., and BORDES, A., “Memory networks,” *CoRR*, vol. abs/1410.3916, 2014.
- [476] WILLIAMS, R. J., “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

- [477] WINOGRAD, T., “Understanding natural language,” in *Cognitive Psychology*, 1972. xiv, 90, 91
- [478] WOLF, F., POGGIO, T., and SINHA, P., “. human document classification using bags of words,” *MIT Tech Report*, 2006.
- [479] WU, J., LU, E., KOHLI, P., FREEMAN, B., and TENENBAUM, J., “Learning to see physics via visual de-animation,” pp. 153–164, 2017.
- [480] WU, Q., WANG, P., SHEN, C., VAN DEN HENGEL, A., and DICK, A. R., “Ask me anything: Free-form visual question answering based on knowledge from external sources,” in *CVPR*, 2016. 3, 44
- [481] XIAO, J., HAYS, J., EHINGER, K., OLIVA, A., and TORRALBA, A., “Sun database: Large-scale scene recognition from Abbey to Zoo,” 2010.
- [482] XIAO, J., HAYS, J., EHINGER, K. A., OLIVA, A., and TORRALBA, A., “Sun database: Large-scale scene recognition from abbey to zoo,” *IEEE*, 2010.
- [483] XIONG, C., MERITY, S., and SOCHER, R., “Dynamic memory networks for visual and textual question answering,” in *ICML*, 2016. 3, 44
- [484] XU, H. and SAENKO, K., “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” *arXiv:1511.05234*, 2015.
- [485] XU, H. and SAENKO, K., “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in *ECCV*, 2016. 3, 44
- [486] XU, N., PRICE, B. L., COHEN, S., YANG, J., and HUANG, T. S., “Deep interactive object selection,” in *CVPR*, 2016.
- [487] YADOLLAHPOUR, P., BATRA, D., and SHAKHNAROVICH, G., “Discriminative re-ranking of diverse segmentations,” in *CVPR*, 2013.
- [488] YAMAGUCHI, K., STRATOS, K., SOOD, A., MITCHELL, M., MENSCH, A., GOYAL, A., HAN, X., DODGE, J., DAUME, H., BERG, T. L., and BERG, A. C., “Understanding and predicting importance in images,” 2012.
- [489] YANG, J., PRICE, B., COHEN, S., LEE, H., and YANG, M.-H., “Object contour detection with a fully convolutional encoder-decoder network,” in *CVPR*, 2016.
- [490] YANG, Y. and RAMANAN, D., “Articulated pose estimation with flexible mixtures-of-parts,” 2011.
- [491] YANG, Y., TEO, C., DAUMÉ III, H., and ALOIMONOS, Y., “Corpus-guided sentence generation of natural images,” in *EMNLP*, 2011.
- [492] YANG, Y., BAKER, S., KANNAN, A., and RAMANAN, D., “Recognizing proxemics in personal photos,” in *CVPR*, IEEE, 2012.

- [493] YANG, Z., HE, X., GAO, J., DENG, L., and SMOLA, A., “Stacked attention networks for image question answering,” 2015. 80, 82, 83
- [494] YANG, Z., HE, X., GAO, J., DENG, L., and SMOLA, A., “Stacked attention networks for image question answering,” 2016.
- [495] YANG, Z., HE, X., GAO, J., DENG, L., and SMOLA, A. J., “Stacked attention networks for image question answering,” in *CVPR*, 2016. xi, 3, 4, 5, 10, 44, 55, 58, 59, 61, 63, 66, 126, 127, 129
- [496] YAO, B. and FEI-FEI, L., “Modeling mutual context of object and human pose in human-object interaction activities,” 2010.
- [497] YI, K., WU, J., GAN, C., TORRALBA, A., KOHLI, P., and TENENBAUM, J. B., “Neural-symbolic vqa: Disentangling reasoning from vision and language understanding,” *arXiv preprint arXiv:1810.02338*, 2018.
- [498] YOONESSI, A. and BAKER, C. L., “Contribution of motion parallax to segmentation and depth perception,” *Journal of Vision*, vol. 11, no. 9, 2011.
- [499] YU, L., PARK, E., BERG, A. C., and BERG, T. L., “Visual madlibs: Fill-in-the-blank description generation and question answering,” in *ICCV*, 2015. 9
- [500] YU, X. and ALOIMONOS, Y., “Attribute-based transfer learning for object categorization with zero or one training example,” 2010.
- [501] ZAIDAN, O., EISNER, J., and PIATKO, C., “Using annotator rationales to improve machine learning for text categorization,” in *NAACL - HLT*, 2007.
- [502] ZAVESKY, E. and CHANG, S.-F., “CuZero: Embracing the frontier of interactive visual search for informed users,” in *Proceedings of ACM Multimedia Information Retrieval*, 2008.
- [503] ZAVESKY, E. and CHANG, S.-F., “Cu-zero: Embracing the frontier of interactive visual search for informed users,” in *MIR*, 2008.
- [504] ZHANG, H., XU, T., LI, H., ZHANG, S., HUANG, X., WANG, X., and METAXAS, D., “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” 2017. 13
- [505] ZHANG, P., GOYAL, Y., SUMMERS-STAY, D., BATRA, D., and PARIKH, D., “Yin and yang: Balancing and answering binary visual questions,” *CoRR*, vol. abs/1511.05099, 2015. 4, 11, 35, 52, 54, 55
- [506] ZHANG, P., GOYAL, Y., SUMMERS-STAY, D., BATRA, D., and PARIKH, D., “Yin and yang: Balancing and answering binary visual questions,” in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 5014–5022, IEEE, 2016.

- [507] ZHANG, W., YU, S. X., and TENG, S.-H., “Power svm: Generalization with exemplar classification uncertainty,” in *CVPR*, 2012.
- [508] ZHAO, J., WANG, T., YATSKAR, M., ORDONEZ, V., and CHANG, K.-W., “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” 2017.
- [509] ZHOU, B., TIAN, Y., SUKHBAATAR, S., SZLAM, A., and FERGUS, R., “Simple baseline for visual question answering,” *CoRR*, vol. abs/1512.02167, 2015.
- [510] ZHOU, X. S. and HUANG, T. S., “Relevance feedback in image retrieval: A comprehensive review,” *Proceedings of ACM Multimedia Systems*, 2003. 3, 44
- [511] ZHU, Y., GROTH, O., BERNSTEIN, M., and FEI-FEI, L., “Visual7w: Grounded question answering in images,” in *CVPR*, 2016.
- [512] ZIEN, A. and ONG, C. S., “Multiclass multiple kernel learning,” 2007.
- [513] ZITNICK, C. L. and PARIKH, D., “The role of image understanding in contour detection,” 2012.
- [514] ZITNICK, C. L., AGRAWAL, A., ANTOL, S., MITCHELL, M., BATRA, D., and PARIKH, D., “Measuring machine intelligence through visual question answering,” *AI Magazine*, vol. 37, no. 1, pp. 63–72, 2016.
- [515] ZITNICK, C. L., AGRAWAL, A., ANTOL, S., MITCHELL, M., BATRA, D., and PARIKH, D., “Measuring machine intelligence through visual question answering,” *AI Magazine*, vol. 37, no. 1, pp. 63–72, 2016.
- [516] ZITNICK, C. L. and PARIKH, D., “Bringing Semantics Into Focus Using Visual Abstraction,” in *CVPR*, 2013. 2, 16, 18
- [517] ZITNICK, C. L., PARIKH, D., and VANDERWENDE, L., “Learning the Visual Interpretation of Sentences,” in *ICCV*, 2013. 18
- [518] ZITNICK, C. L., VEDANTAM, R., and PARIKH, D., “Adopting Abstract Images for Semantic Scene Understanding,” *PAMI*, 2015. 18
- [519] ZWEIG, G. and BURGESS, C. J., “The microsoft research sentence completion challenge,” Tech. Rep. MSR-TR-2011-129, Microsoft Research, December 2011.